REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
May 2003

3. REPORT TYPE AND DATES COVERED
Annual (1 May 02 - 30 Apr 03)

4. TITLE AND SUBTITLE
Digital Mammography: Advanced Computer-Aided Breast Cancer Diagnosis

5. FUNDING NUMBERS
DAMD17-02-1-0214

6. AUTHOR(S)
Heang Ping Chan Ph.D.

20040223 094

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Ann Arbor, Michigan 48109-1274

E-Mail: chanhp@umich.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The goal of the project is to develop a computer-aided diagnosis (CAD) system for full field digital mammography (FFDM) using advanced computer vision techniques and image information fusion from multiple views and bilateral mammograms to improve lesion detection and characterization. When fully developed, the CAD system can assist radiologists in mammographic interpretation.

During this project year, we have performed the following tasks: (1) Collected a data set of FFDM cases that contain mammographic lesions. (2) Developed a database management program to store the coded case information to facilitate archiving and retrieval of the cases. (3) Deceloping a Laplacian pyramid image enhancement technique for preprocessing the raw image from the FFDM system, which is then as the input to our CAD system. (4) Compared the mass detection accuracy of our CAD algorithm when applied to our processed images and the GE's processed images and found that the detection sensitivities on the two sets of images were within a few percent over the entire FP range of interest. (5) Compared the performance of the mass detection system on digitized film mammograms and DMs, and found that, with adjustment of the processing parameters in the algorithm, the detection accuracies of the algorithms on both sets of images are comparable. (6) Investigated the effects of the image enhancement filter on the accuracy of mass detection and found that, by replacing the DWCE filter with an adaptive ring filter, the sensitivity of mass detection can be improved up to 15 %. (7) Compared the mammographic density segmented from digitized film mammograms and DMs from the same patients, and found that the correlation of the segmented breast density between the two types of images is very high but the estimated percent dense area on DMs is, on average, about 5% lower than that estimated from digitized film mammograms. (8) Developing two-view information fusion method for correlating the detected lesions on the CC- and MLO-view mammograms, and found that the detection accuracy for microcalcifications can be improved by fusing of information from the two mammographic views.

In conclusion, we have investigated a number of CAD techniques for detection of masses and microcalcifications on mammograms. We have made progress in the six tasks proposed in the project. This lays the strong foundation for us to continue the development of the CAD system for digital mammograms in the coming years.

14. SUBJECT TERMS
Breast Cancer
full field digital mammography, computer-aided diagnosis, breast cancer diagnosis

15. NUMBER OF PAGES
44

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unlimited

Award Number: DAMD17-02-1-0214

TITLE: Digital Mammography: Advanced Computer-Aided Breast Cancer Diagnosis

PRINCIPAL INVESTIGATOR: Heang Ping Chan Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, Michigan 48109-1274

REPORT DATE: May 2003

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

Table of Contents

(4) Introduction

Computer-aided diagnosis (CAD) has been shown to be useful as a second opinion to radiologists for breast cancer detection on mammograms. All current CAD systems have been developed for digitized screen-film mammograms. With the recent advent of full field digital mammography (FFDM) systems, it is important to develop CAD systems specifically designed for direct digital mammograms (DMs) in order to fully exploit the advantages of FFDM. Although many computer vision techniques developed for digitized films may be used for DMs, proper adaptation and extensive training of the current algorithms for the new type of images will be required. More importantly, new techniques still need to be developed to further improve the current algorithms.

The goal of the proposed research is to develop a CAD system for FFDM using advanced computer vision techniques. The proposed CAD system will assist radiologists with detection and classification of breast lesions. Previous CAD methods for lesion detection and characterization are generally based on image features extracted from a single view. Our proposed approach is distinctly different from the previous approaches in that innovative techniques will be developed to fuse image information from multiple views and bilateral mammograms to improve lesion detection and characterization. We hypothesize that these advanced intelligent techniques will lead to an effective CAD system for FFDM.

The following specific aims will be addressed: (1) Collection of a database of DMs and design of a database management system. (2) Development of single-view computer vision techniques for mass detection and classification in DMs. (3) Development of single-view computer vision techniques for microcalcification detection and classification in DMs. (4) Development of methods for correlation of image information from two-view mammograms. (5) Development of methods for correlation of image information from bilateral mammograms. (6) Comparison of the detection and classification accuracy of the multiple-image fusion CAD system for DMs with the performance of the one-view CAD system and other CAD systems by receiver operating characteristic (ROC) analysis.

We will first adapt our current algorithms for digitized mammograms to DMs, taking into account the differences in the imaging characteristics between DMs and digitized film mammograms. In addition, new computer vision techniques will be developed in each of the four areas to improve the current methods and to exploit the higher contrast sensitivity and higher detective quantum efficiency of digital mammography detectors over screen-film systems. We will develop novel regional registration methods for identifying corresponding lesions on craniocaudal (CC) and mediolateral oblique (MLO) views and comparing bilateral mammograms. The multiple image information will be fused with fuzzy classification to reduce false positives and to improve lesion detection sensitivity. Multiple-view features of a lesion will be merged using neural networks or other classifiers for classification of malignant and benign lesions. A large database of DMs will be collected from our patient population and extensive training and independent testing of the new CAD system will be performed. The test performance of the advanced multiple-image fusion CAD algorithms for detection and characterization of lesions on DMs will be compared with the one-view approach on DMs as well as the performances of CAD systems for digitized film mammograms using ROC methodology.

Digital mammography not only has the potential to detect breast cancer in an earlier stage, it will also facilitate consultation via teleradiology in remote or rural regions where expert mammographers may not be readily available. An effective CAD system will be particularly useful for providing an additional on-site or remote second opinion. This will be highly relevant to women in the military, especially when they are stationed in remote areas. FFDM in combination with CAD will fully utilize the potential of digital mammography to improve the health care of women both in the military and in the general population.

(5) Body

This is the final report of our project. In the project period (4/20/98-4/19/03), we have performed a number of studies to develop the digital stereoscopic imaging technique. The detailed studies and results have been reported previously in the annual progress reports. A summary of some of the important accomplishments follows.

(A) Collection of database

We have been collecting a database of digital mammograms (DMs) with mammographic masses or microcalcifications for the development of our computer-aided diagnosis (CAD) algorithms. We have collected about 80 cases. The patients were diagnosed with abnormalities in their mammograms during their normal clinical care, either by routine screening or by referral to our breast imaging clinic for evaluation. The digital mammograms were acquired with a GE Senographe 2000D full field digital mammography (FFDM) system. The system has a flat panel detector consisting of amorphous Si active matrix with CsI phosphor. The pixel size of the system is 100 μm X 100 μm. The gray level resolution of the system is 14 bits for the raw images and 12 bits for the processed images. After acquisition, the digital image files are transmitted to the Siemens Archive which is the PACS system used in our department for storage of all clinical digital images.

With Institutional Review Board (IRB) approval, we retrieved the digital mammograms from the Siemens Archive to our laboratory. We have developed a database management program based on Microsoft Access to process the images downloaded to our system. For each mammogram file, all patient identifiers are first removed from the image header. The patient name is replaced with a code number. The image is then named by the code number, the view (craniocaudal, mediolateral oblique, or mediolateral), and the exam year. A record is also generated in the database file for each image. The record keeps the code number, the lesion type, the view, and the exam date information for each case. When the pathology of the case is available, the malignant or benign information of the lesion is also entered. Each case in the database is read by an experienced MQSA radiologist to mark the lesion location. For microcalcification cases, the radiologist measures the diameter of the cluster, and provides description of its distribution, morphology, and visibility of the microcalcifications. For mass cases, the radiologist measures the diameter of the mass, and provides description of its margin, shape, spiculated or non-spiculated, the visibility, and the density of the mass relative to that of the parenchyma. For all cases, the radiologist also provides BI-RADS description of the breast density and estimates the likelihood of malignancy of the lesion. These descriptions are entered into the database for each case as a reference for future analysis.

(B) Pre-processing technique for digital mammograms

DMs generally are pre-processed with proprietary methods by the manufacturer of the FFDM system before being displayed to readers. The image pre-processing method used depends on the manufacturer of the FFDM system. In an effort to develop a CAD system that is less dependent on the FFDM manufacturer's proprietary preprocessing methods, we use the raw FFDM as input to our CAD system. We are developing a multi-scale preprocessing scheme for image enhancement.

Methods:

Multi-scale methods have been used for contrast enhancement of medical images recently. Since a multi-scale method uses the information from a large number of frequency channels extracted from the image adaptively, it is more flexible and versatile than the commonly used enhancement methods, such as unsharp masking, which uses a small number of frequency channels. Two types of multi-scale methods have been used as the preprocessing methods for the contrast enhancement of mammograms: the wavelet method and the Laplacian pyramid method. A previous study has shown that, for image enhancement purpose, using a Laplacian pyramid method has more advantages than using the fast wavelet transformation (1). In this project, therefore, we chose the Laplacian pyramid method as our preprocessing method.

A flow-chart of our preprocessing method is shown in Fig. 1. In brief, the breast region is first segmented automatically by using Otsu's method into the background and the breast region. Second, the Laplacian pyramid method is used to decompose the breast image into multi-scales. A nonlinear weight function based on the pixel gray level from each of the low-pass components is designed to enhance the high-pass components. The details are described below.

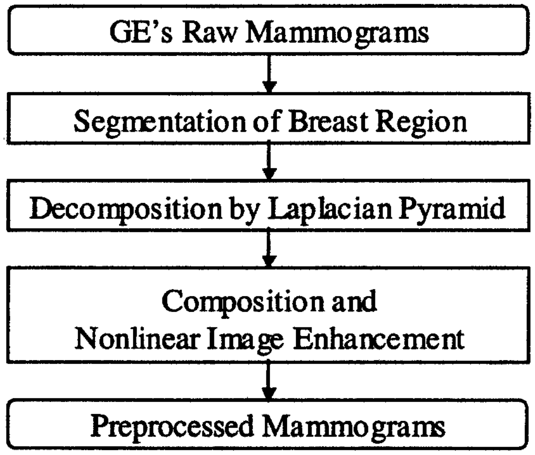

GE's Raw Mammograms
Segmentation of Breast Region
Decomposition by Laplacian Pyramid
Composition and
Nonlinear Image Enhancement
Preprocessed Mammograms

Fig. 1. Flowchart of the method for pre-processing of the raw digital mammogram before the CAD algorithm.

(a) Segmentation of Breast Region

A two-step algorithm was developed for the segmentation of breast region. First, Otsu's method is used to calculate a threshold and binarize the original image. Second, a labeling method using 8-connectivity is used to identify the connected regions on the binary image. The region with the largest area will be considered to be the breast region.

(b) Decomposition by Laplacian pyramid

The Laplacian pyramid decomposition is a multi-scale method that was first introduced as an image compression technique (2). The general scheme is shown in Fig. 2. The Laplacian pyramid is a sequence of error image $L_0, L_1, \cdots, L_n$. Each is the difference between two levels of the Gaussian pyramid. The decomposition of the image from level *l* to level *l+1* can be expressed mathematically as in the following equation:

$$L_l = g_l - Expand(g_{l+1}) \tag{1}$$

where

$$Expand(g_{l+1}) = 4\sum_{m=-2}^{2}\sum_{n=-2}^{2} w(m,n)\cdot g_l(\frac{i-m}{2},\frac{j-n}{2}) \tag{2}$$

$$g_{l+1}(i,j) = \sum_{m=-2}^{2}\sum_{n=-2}^{2} w(m,n) g_l(2i+m,2j+n) \tag{3}$$

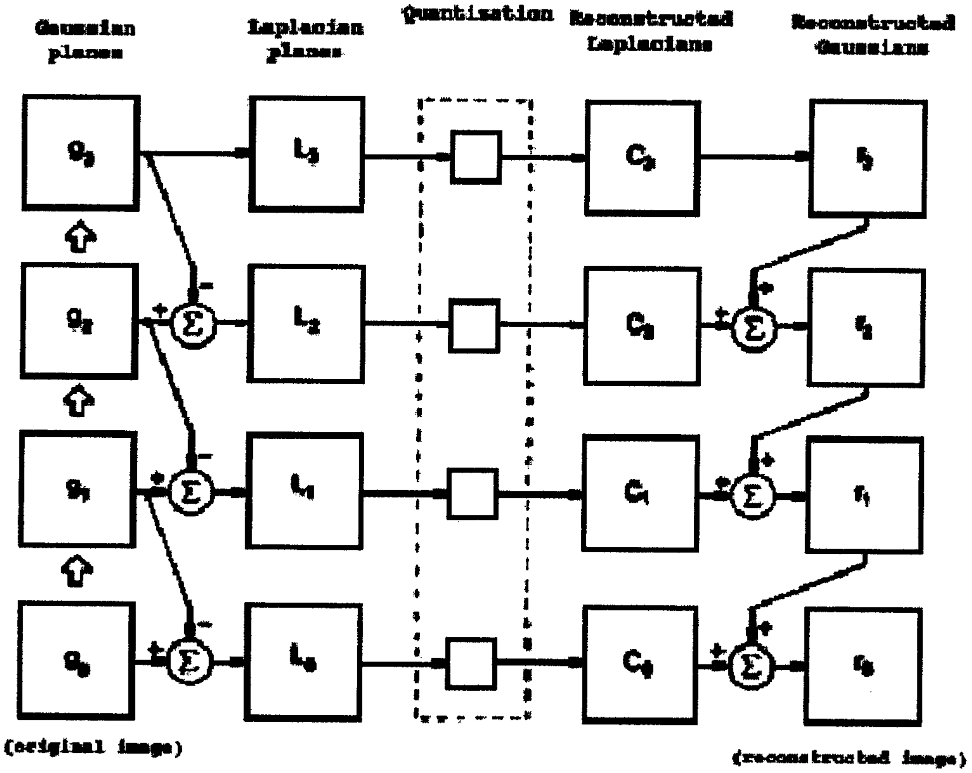


Fig. 2. Laplacian pyramid decomposition and reconstruction of an image.

(c) Reconstruction and nonlinear image enhancement

The original image can be recovered by summing all the levels of the Laplacian pyramid. For the purpose of image enhancement, the image at each level of the Laplacian pyramid that corresponds to a bandpass image is mapped by a nonlinear function. Several types of nonlinear function have been examined for enhancement of x-ray images. In ref. (3), a power law with a linear lower and upper cutoff was used. Stahl et al (4) used a power law bounded by linear functions for very small and very large contrast.

In this study, we used a nonlinear function that incorporates the information from each bandpass image. The defined nonlinear function is given by

$$r(l) = \alpha \cdot Expand(g_{l+1}) + \beta \cdot (Expand(g_{l+1}))^{p} \cdot L_l \tag{4}$$

where α, β, and p are the constant values experimentally chosen for each frequency level.

Results:

A set of digital mammograms consisting of the craniocaudal (CC) view and the mediolateral oblique (MLO) view of both breasts of the patient were used for training and testing the preprocessing method.

Fig. 3 shows an example of the segmentation of breast region. Fig. 4 shows the enhancement results of Fig. 3 obtained by using different mapping functions. The processed image enhanced by the GE proprietary method is also shown for comparison. The results obtained with the enhancement functions in ref. (1) and ref. (2) showed over-enhancement inside the breast and under-enhancement around the breast boundary. The GE processed image and our processed image have similar appearance.

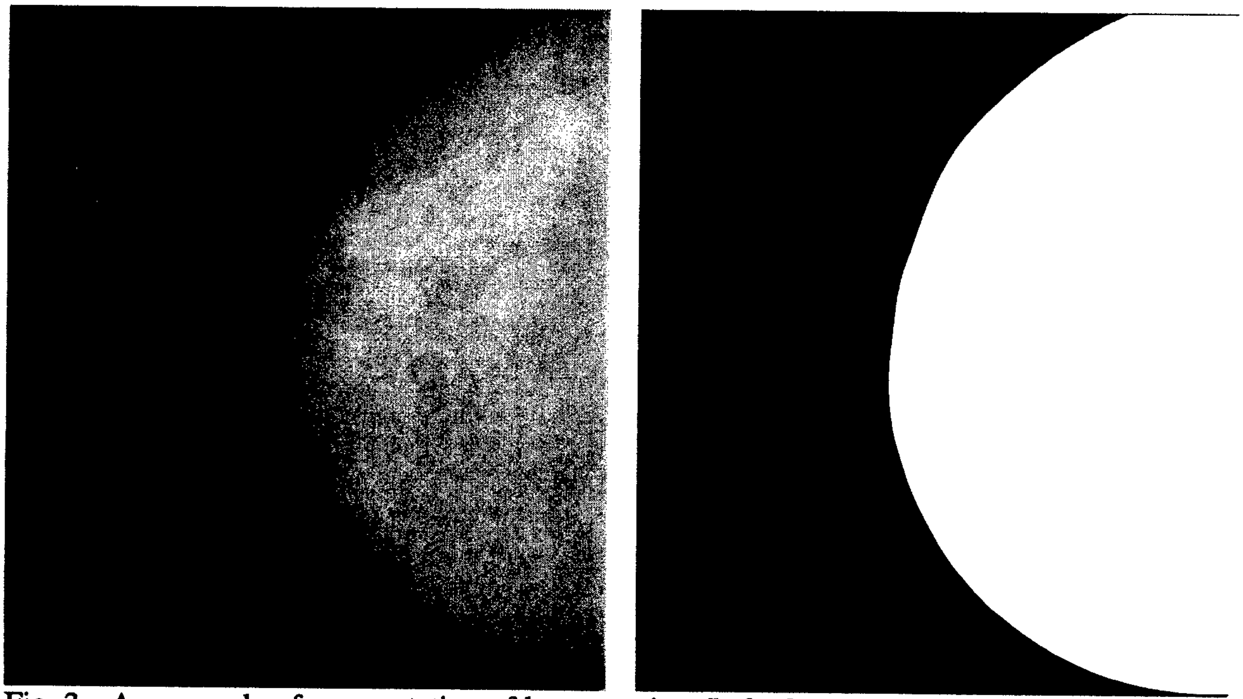


Fig. 3 An example of segmentation of breast region. Left: GE raw image. Right: the result of segmentation of breast region.

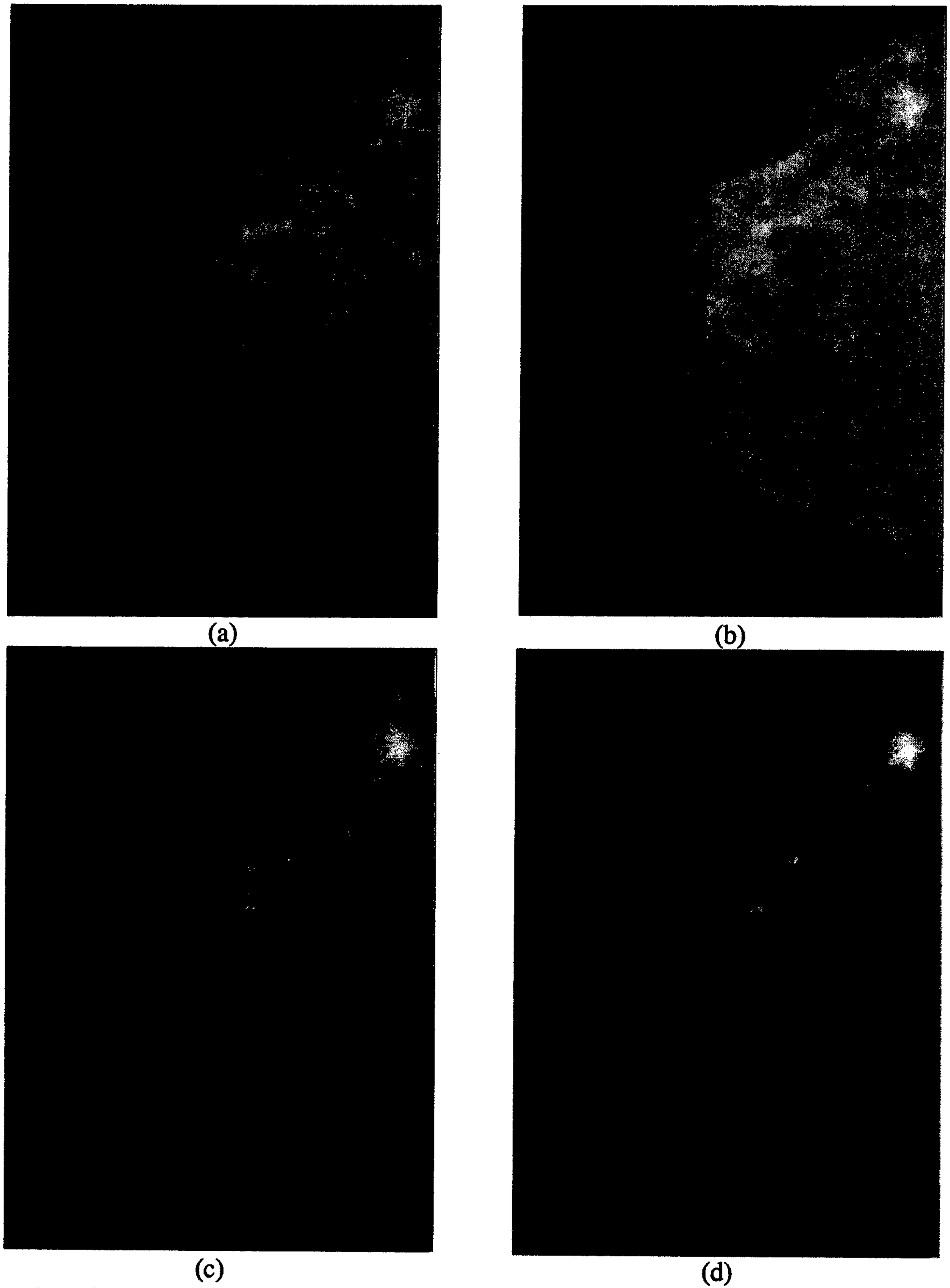

(a) (b) (c) (d)

Fig. 4. The results of enhancement by using different mapping functions. (a) Using enhancement function in ref (1). (b) Using enhancement function in ref. (2). (c) our processed image. (d) GE processed image.

(C) Comparison of Laplacian preprocessing method with GE method by evaluation of mass detection accuracy

For comparison of the GE processed image and our processed image, we analyzed the relative performance of our CAD scheme on these two types of images.

Methods:

Our CAD system consisted of four steps. The input mammogram is first processed with an adaptive density-weight contrast enhancement (DWCE) filter followed by clustering-based region growing to identify suspicious breast structures. Each of these structures is processed by a local refinement stage. Morphological and texture features are then extracted from the refined objects. Rule-based and linear classifiers have been trained with digitized mammograms previously to differentiate masses from normal tissues. In this pilot study, we evaluated the differences in detection accuracy between digital mammograms preprocessed by the GE method and by our Laplacian pyramid method as a measure of the image information on the images.

A data set of 58 digital mammograms was randomly selected from our database. The GE processed images are the images directly output by the FFDM system for display to the radiologists for interpretation. The raw digital images are usually not read by radiologists. Our GE FFDM system was set up so that the both the GE processed images and the raw images are sent from the system and stored in the Siemens Archive. We downloaded the raw images for each case to our laboratory for preprocessing with our Laplacian pyramid method as described above. The mass detection system that was trained with digitized film mammograms in our previous studies was applied to the two sets of images. Since the CAD system performance will likely be degraded similarly for both types of images, the relative performance of the mass detection accuracy may not depend on retraining. Therefore, in this comparison study, we used the CAD system before it was retrained for digital images.

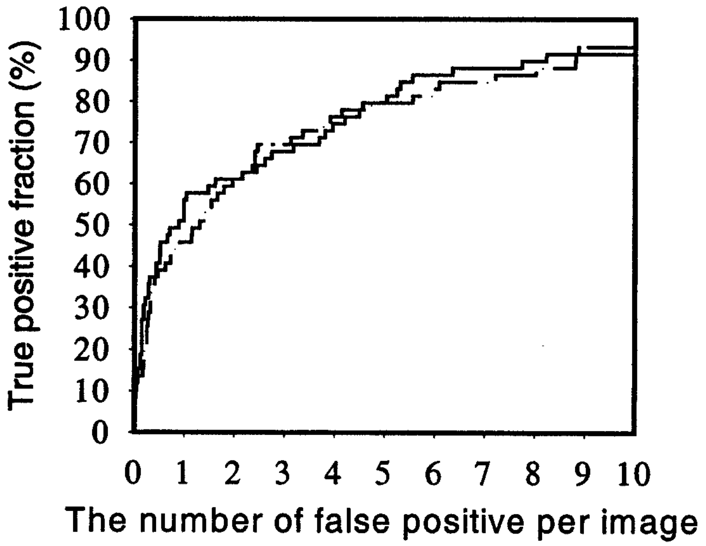


Fig. 5. Comparison of FROC curve for mass detection of mass on GE processed mammograms and our processed mammograms. Solid line: our processed mammograms. Dash-dot line: GE processed mammograms.

Results and Conclusion:

The detection accuracy of the CAD system on the digital mammograms is evaluated by free-response receiver operating characteristic (FROC) analysis. Fig. 5 shows the comparison of FROC curves for detection of masses on the GE images and on our processed images. The sensitivity of the mass detection on the two types of processed images is within a few percent over the entire FP range. This result confirms that the information available for mass detection is similar in the two types of processed images.

After this baseline is established, we can adjust the preprocessing parameters to optimize the CAD system without depending on the manufacturer's algorithm. We believe that this will allow us to develop a CAD system that can be adapted to other FFDM systems easily.

(D) Comparison of density segmentation on digitized screen-film mammograms and digital mammograms

Previous studies have found that there is a strong correlation between mammographic breast density and the risk of breast cancer. Mammographic breast density has been used by researchers in many studies to estimate breast cancer risk of epidemiological factors, monitor the effects of preventive treatments such as tamoxifen or dietary interventions, monitor the breast cancer risk of hormone replacement therapy, and investigate factors affecting mammographic sensitivity and cancer prognosis. Digital mammographic systems have recently been introduced into clinical use. In this study, we compared the breast density estimated on pairs of digital mammogram (DM) and digitized screen-film mammogram (SFM) obtained from the same patients.

Methods:

We are comparing image information on DMs and SFMs for radiologist's interpretation and computerized image analysis. One hundred forty-five pairs of DM and SFM (76 CC views and 69 MLO views) were collected with IRB approval from 68 patients. The time interval between the DM and SFM ranged from 0 to 118 days (median=21 days). The SFMs were acquired with GE DMR systems and the DMs were acquired with the GE Senographe 2000D system. Both the DMs and the SFMs were acquired with automated exposure techniques that selected the appropriate target, filter, and kVp. The SFMs were digitized with a laser film scanner. The breast boundaries on the DMs and SFMs were detected automatically by the computer. The mammograms were displayed on a workstation with a graphical user interface (see Fig. 6) that allowed interactive thresholding of the gray level histograms to segment the dense region from the fatty region. The DMs and SFMs were segmented independently in separate sessions so that the observer could not compare the density of the corresponding DM and SFM. Hard copies of the displayed images were available for reference during segmentation. The mammographic density was estimated as the percent dense area relative to the breast area, excluding the pectoral muscle in the MLO views.

Results:

An example of density segmentation with interactive thresholding using the graphical user interface is shown in Fig. 6. Fig. 7 shows the comparison of the percentage area of breast density on both digitized SFMs and DMs by an observer.

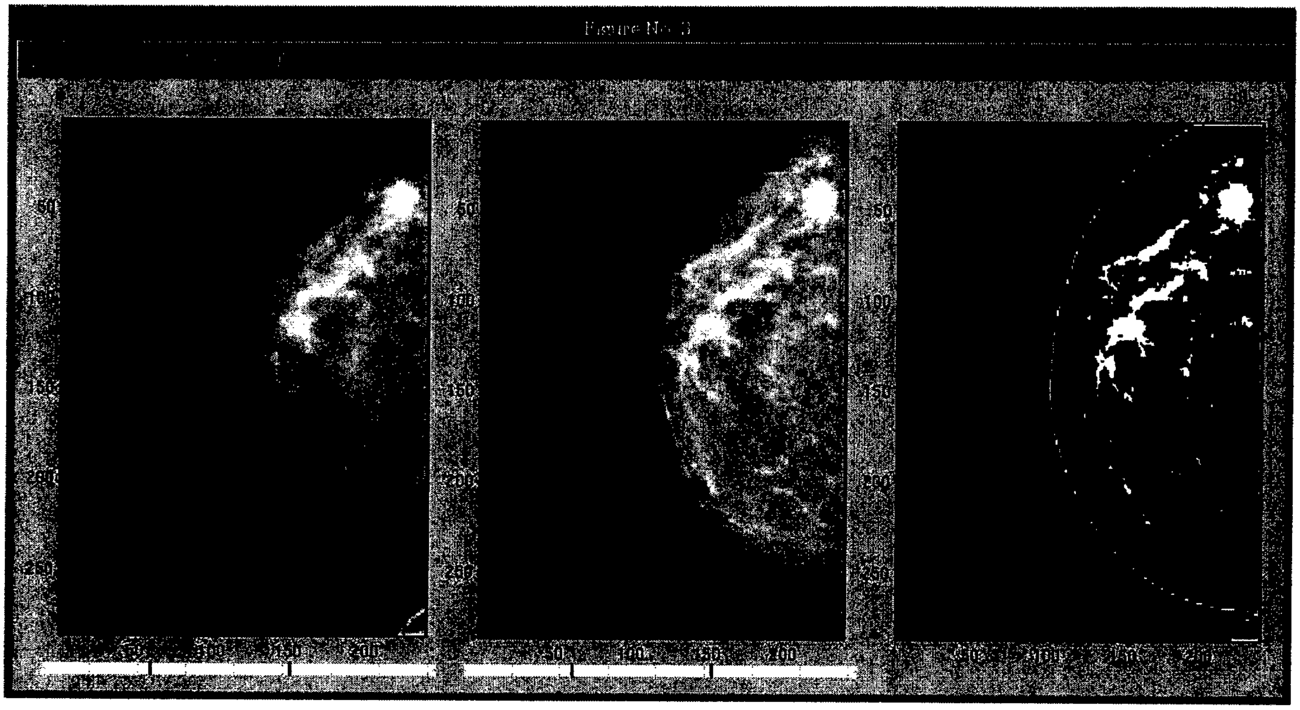


Fig. 6. An example of density segmentation on digital mammograms. Left: GE processed image. Middle: Our processed image. Right: segmentation result of breast density.

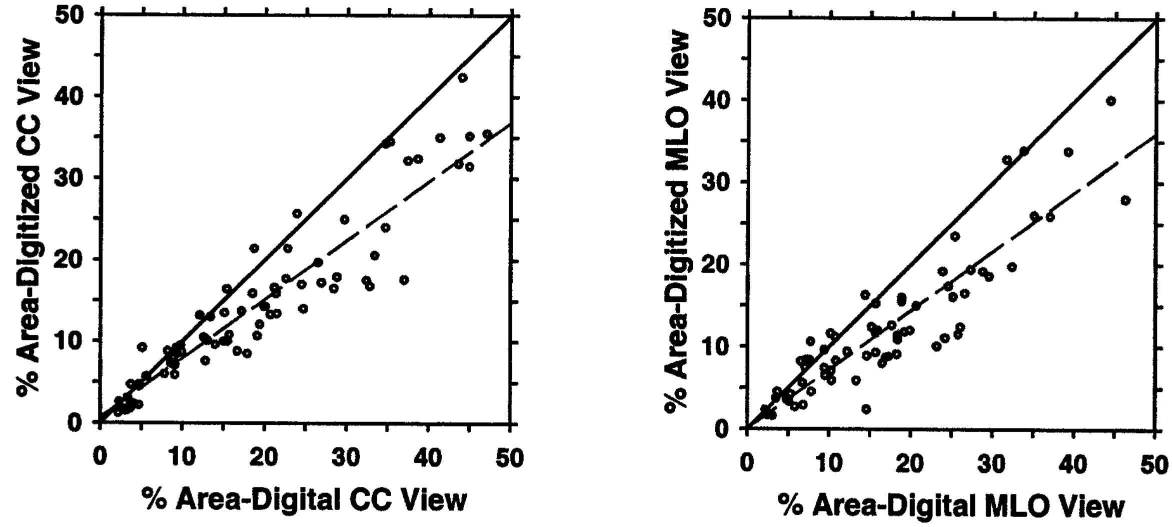


Fig. 7. Comparison of the percent mammographic density obtained on digital and digitized mammograms. Left: CC view, correlation coefficient = 0.94; Right: MLO view, correlation coefficient = 0.92. Dash line: linear regression of the data; solid line: diagonal.

The correlation between the mammographic density on SFM and DM was 0.94 and 0.92, the root-mean-square residual was 4.5% and 4.6%, and the average ratio of mammographic density estimated on SFM to that on DM of the same breast was 1.18 and 1.22, respectively, for CC and MLO views. The differences in the percent dense area between the DM and SFM were statistically significant (paired t test: p<0.0000001) for both views. The DMs used harder beams (Mo/Mo 4.5%, Mo/Rh 22.4%, Rh/Rh 73.1%) while the SFMs used softer beams (Mo/Mo 44.2%, Mo/Rh 48.1%, Rh/Rh 7.8%). The peak potential used for DM was 1 to 5 kVp higher than that for SFM in 84% of cases.

Conclusion:

Breast density on DMs generally appears to be lower than that on SFMs because of the harder beam quality used and image processing applied to the DMs. The lower density may improve the mammographic sensitivity for lesion detection on dense breasts. However, for patients with SFMs and DMs taken over time, comparison of serial mammograms for breast density changes will be problematic.

(E) Computer aided diagnosis system for mass detection: comparison of performance on digital mammograms and digitized mammograms

A CAD system for the detection of masses on digitized screen-film mammograms (DFMs) was developed in our previous studies. We are developing a mass detection system for mammograms acquired directly by a FFDM system. In this study, we compared the performance of the two systems on pairs of DM and DFM images obtained from the same patients.

Methods:

As discussed above, our CAD system consisted of four steps: processing with an adaptive DWCE filter, clustering-based region growing and local refinement, extraction of morphological and texture features, and rule-based and linear classification. In this study, the mass detection system was adapted to DMs by retraining. A data set of 65 cases containing 135 DMs acquired with a GE FFDM system and the 135 DFMs of the same view for the same breast was used. The time interval between the DFM and the corresponding DM was 0 to 118 days. The data set contained 69 masses. The true locations of the masses were identified by an experienced radiologist. The CAD system trained with screen-film system was applied to the set of DFM images. For the DM images, we preprocessed the raw images with the Laplacian pyramid technique, described above. The CAD system parameters were retrained and then applied to the DM images.

Results:

With initial retraining of the CAD system, our mass detection scheme could perform equally well on the DFMs and the DMs. Fig. 8 shows the comparison of case-based FROC curves for mass detection on the DFMs and DMs. The FROC curves achieved a sensitivity of 80% at 2.1 FP marks/image for both the DM and DFM images. The difference in sensitivity between the two types of images was less than 10% over the entire FP ranges shown.

Conclusion:

With retraining, our mass detection CAD scheme can be useful for detecting masses on both DMs and DFMs. Further study is underway to improve the various stages of the mass detection system based on the properties of the DM images.

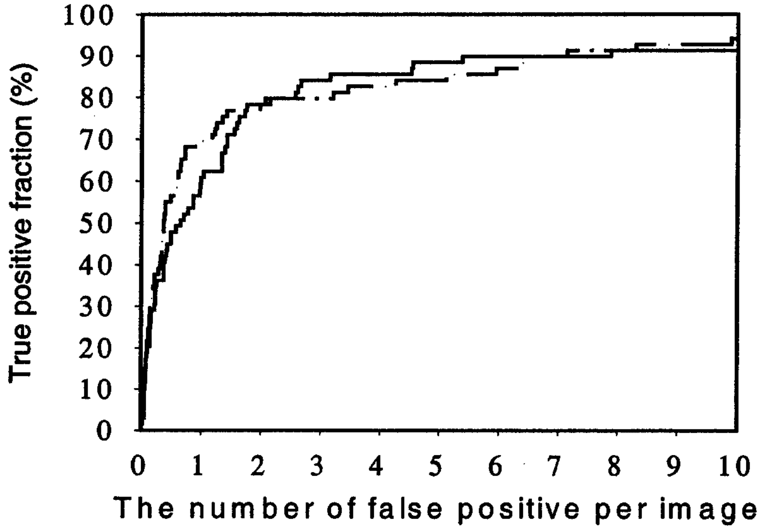


Fig. 8. Comparison of case-based FROC curves for mass detection on the DFM and DM images. Solid line: DMs. Dash-dot line: DFMs.

(F) Adaptive ring filter for enhancement of masses

As discussed above, our CAD system consisted of four steps: enhancing masses or breast structures using an image enhancement filter, clustering-based region growing and local refinement, extraction of morphological and texture features, and rule-based and linear classification. The first mass enhancement step is an important step because it accentuates the mass or other breast structures on an input image to facilitate screening the image for candidate lesions. If the lesion is missed in this step, it will not be recovered in later steps. For our mass detection program previously developed for digitized screen-film mammograms, we developed a density weighted contrast enhancement (DWCE) filter for image enhancement. In the current development of a mass detection system for direct digital mammograms (DMs), we are investigating the effectiveness of different types of filters for enhancing mammographic masses. In this study, an adaptive ring filter is being compared with the DMCE filter.

Methods:

The adaptive ring filter is a type of convergence index filter that was first introduced for detection of nodules on chest x-rays (5). In brief, the gradient vector directions around each pixel in the breast region are calculated. At a mass region, the pixel values tend to be high near the center of the mass and decrease as the distance from the center increases, similar to the height around the peak of a mountain. The gradient vector directions around the mass therefore tend to converge towards the center of the mass. After filtering by the ring filter, the convergence points in the breast region can be considered to be the candidate regions of masses. We chose the top 20 local maximum points in the breast image as the candidates. The corresponding objects were obtained by using a clustering method. A flow-chart was shown in Fig. 9.

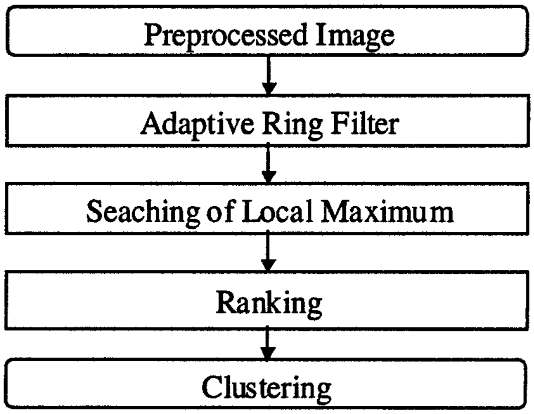
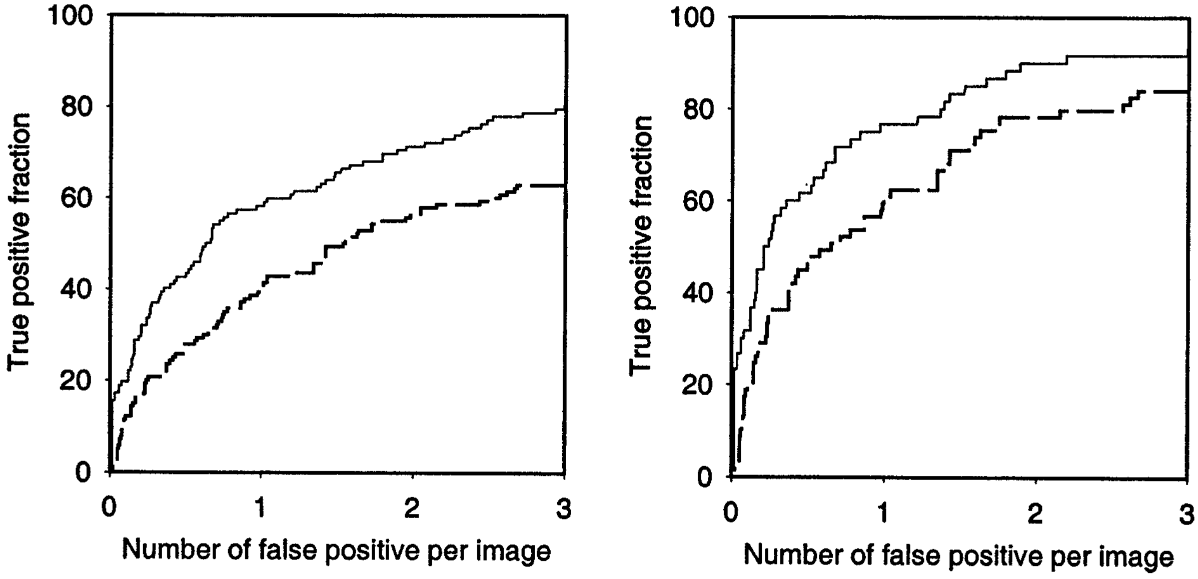


Fig. 9. The flow-chart for detection of candidates of masses

A data set of 65 cases containing 135 DMs acquired with a GE FFDM system as described in Section (E) was used in this comparison study. The DM images were processed in two ways: one is that the DWCE filter was used to enhance the images and thus to facilitate the detection of mass candidates, the other way is that the DWCE filter was replaced by the adaptive ring. For both filters, the next three steps of the mass detection algorithm were the same as those discussed above, except that the parameters used in these steps were retrained for DMs separately for use with the adaptive ring filter and the DWCE filter.

Results:



(a)

(b)

Fig. 10. Comparison of adaptive ring filter (solid line) with DWCE filter (dash line). (a) Image-based FROC curve. (b) Case-based FROC curves

The FROC curves for the DWCE filter and the adaptive ring filter are compared in Fig. 10. The sensitivity of mass detection using the adaptive ring filter was improved by up to 15% over the range of FPs shown, either considering the detection by image or by case.

Conclusion:

The adaptive ring filter therefore seems to be more effective than the DWCE filter in enhancing the masses for automated detection. Further study is underway to collect a larger data set, to optimize the parameters in the different stages of the CAD system, and to confirm the performance of the adaptive ring filter.

(G) Joint two-view information for computerized microcalcification detection

The CAD systems developed to date use single-view mammograms for lesion detection. In mammoraphic interpretation, radiologists find it very useful to combine the information from two views (CC and MLO views) to confirm true lesions and to exclude false lesions. In our project we propose to develop methods to correlate the information from two mammographic views so as to improve the detection accuracy of the CAD system. In the following, we discuss a preliminary study that we performed to improve detection of microcalcifications.

Methods:

Figure 11 shows an outline of the joint two-view detection method being developed. Microcalcification cluster candidates are first located using our previously developed single-view lesion candidate detection algorithm. To reduce the false-positives among the detected objects, each object is classified using two different classifiers that work in parallel. The first classifier is similar to the single-view lesion candidate classifier that has been used in our previously developed algorithm. The second classifier, referred to as the correspondence classifier, uses object pairs from the CC and MLO views to characterize whether the object pair consists of two true positives, one on each view. The scores from these two classifiers are merged, and the merged score is used for false-positive reduction.

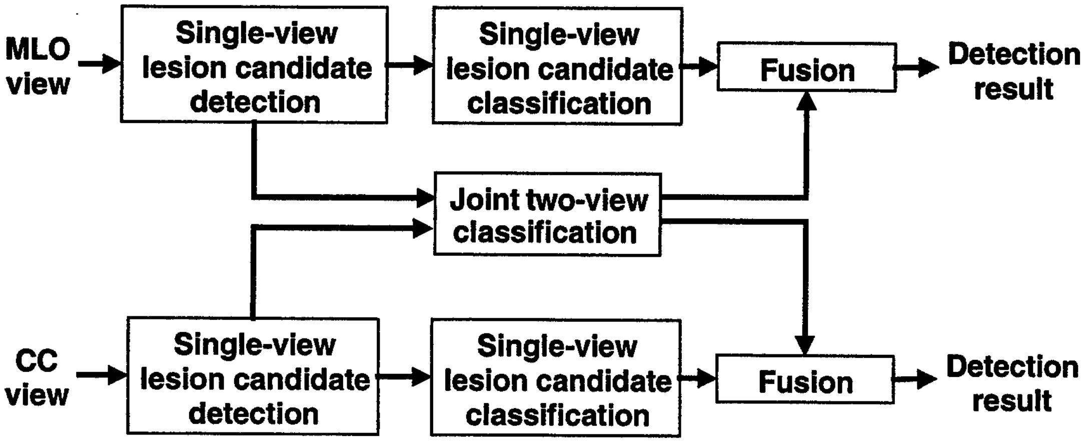

MLO view
Single-view lesion candidate detection
Single-view lesion candidate classification
Fusion
Detection result
Joint two-view classification
CC view
Single-view lesion candidate detection
Single-view lesion candidate classification
Fusion
Detection result

Fig. 11. The block diagram of the computerized joint two-view method for microcalcification detection.

(a) Single-View Lesion Candidate Detection and Classification

The parameters of our single-view lesion candidate detection algorithm were adjusted to provide high-sensitivity, at the cost of a relatively large number of false-positives (FPs). The single-view lesion candidate classification step aims at reducing the FPs while maintaining a high sensitivity. The classifier was designed using stepwise feature selection and linear discriminant analysis (LDA) trained on an independent training set. The feature space used in stepwise selection included the texture and morphological features of the microcalcification clusters, as well as the number of microcalcifications in a cluster and the output scores of a convolution neural network (CNN).

(b) Joint Two-View Classification of True and False Pairs

Joint two-view pair classification distinguishes between true (TP-TP) pairs and false pairs (TP-FP, FP-TP, and FP-FP) by using the similarities between the two objects that constitute the pair. The initial step in this task is to define the object pairs. As described in our previous studies (6) (7), we define the pairs based on the difference between the nipple-to-object distances (NODs) on the CC and MLO views. For each nodule candidate C_{ci} detected on the CC view, the NOD, D_{ci}, is computed, and an annular region of width 2ΔR, centered at the nipple and enclosed between the radii $D_{ci}\pm\Delta R$ is defined on the MLO view. If the centroid of an object C_{Mj} on the MLO view is found inside this annular region, then an object pair (C_{ci}, C_{Mj}) is formed. The width of the annular region, 2ΔR, is determined using training data as 6.0 cm.

A correspondence classifier, based on stepwise feature selection and LDA, is used to estimate the likelihood that the defined pair is a true pair. The feature space used in stepwise selection included the similarity measures of the features that are used in single-view lesion candidate classification.

(c) Information Fusion

The correspondence classifier produces a correspondence score for each object pair. This score is converted into a two-view object score before being combined with the single-view object score. The fusion score for an object is defined as the average of its single- and two-view object scores in this preliminary study.

(d) Data Set

Our training data set consisted of 108 pairs of biopsy-proven CC and MLO mammograms containing microcalcification clusters, collected with IRB approval at the University of Michigan (UM). The mammograms were digitized with a LUMISYS 85 laser scanner at a pixel size of 50 μm x 50 μm and 4096 gray levels. The digitizer was calibrated so that the gray level values are linearly proportional to the optical density (OD) within the range of 0.1 to 4.0 OD units. The mammograms are filtered with a 2x2 box filter and subsampled by a factor of 2 to produce a 0.1 mm x 0.1 mm images prior to processing.

Our independent test set consisted of 116 pairs of mammograms, selected from the University of South Florida (USF) public mammogram database (8). The digitization characteristics of these mammograms were similar to those of the UM database, with the difference that a Lumisys 200 laser scanner was used. The test data set contained 254 microcalcification clusters on 232 mammograms. The FP rate was determined by applying the algorithm to an additional 76 normal mammogram pairs (152 mammograms) from the USF data set.

Results:

The detection accuracy is evaluated using film-based and case-based FROC curves. In film-based analysis, a microcalcification cluster detected on one view but missed on the other view is considered as one TP and one false-negative (FN). This method provides a more conservative sensitivity estimate than a case-based analysis. In case-based analysis, a TP was defined as marking a malignant cluster on at least one view.

The prescreening algorithm detected 89% (226/254) of the clusters with an average of 3.5 FPs/image (539/152) on the normal mammograms. Based on the NOD, a total of 5929 object pairs were defined on the normal test images (51 object pairs/mammogram pair) and 523 object pairs were defined on the abnormal test images (6.8 object pairs/mammogram pair). The object pairs were classified by the correspondence classifier designed using the training set. The pair scores were converted into object scores and were fused with single-view scores. The final FROC curve obtained by the fusion method is compared to the single-view FROC curve in Figure 12 and 13, for film-based scoring and case-based scoring, respectively. The single-view detection algorithm had a film-based sensitivity of 86% at 0.6 FPs/image. At the same sensitivity, the two-view detection algorithm produced 0.4 FPs/image. The sensitivity of the single-view and two-view detection algorithms was 79% and 83%, respectively, at 0.1 FPs/image. If correct detection was defined as marking a malignant cluster on at least one view, the two-view detection algorithm achieved a sensitivity of 90% at 0.1 FPs/image.

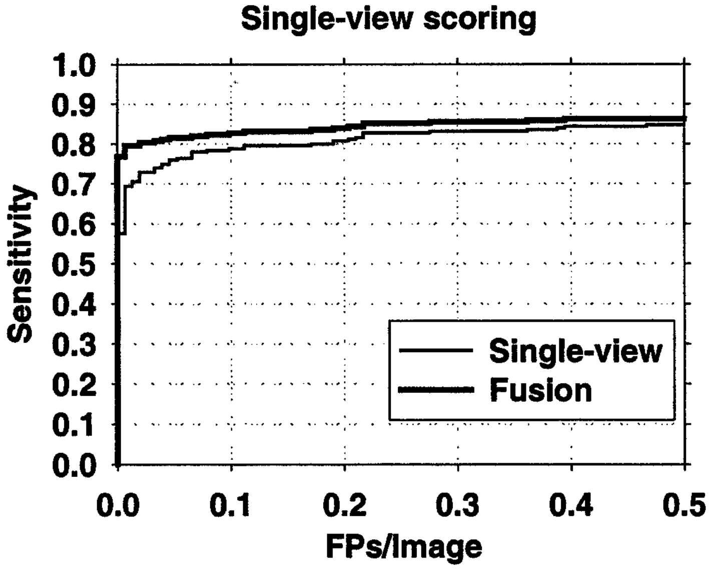


Fig. 12. Comparison of the film-based FROC curves for the single-view detection and fusion methods.

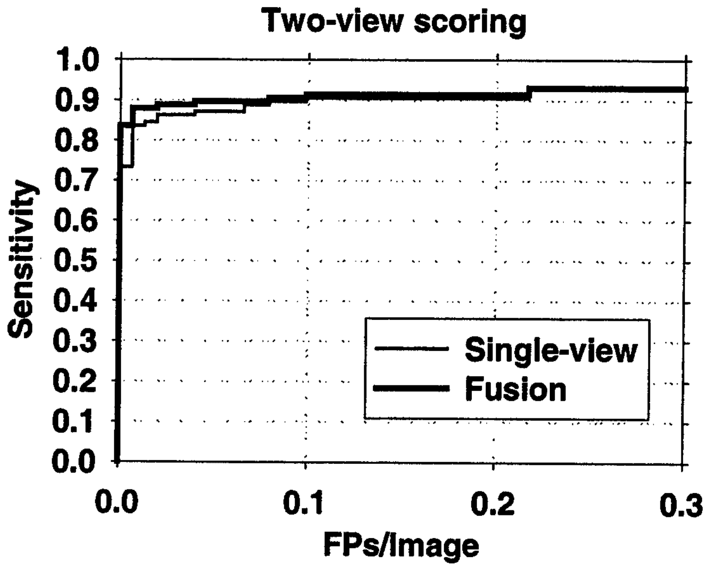


Fig. 13. Comparison of the case-based FROC curves for the single-view detection and fusion methods.

Conclusion:

The correspondence of geometric, morphological, textural and neural network features of cluster candidates on two different views provides valuable information for improving the accuracy of computerized microcalcification detection. Further study is underway to optimize the feature extraction and selection processes, as well as the trainnig of the correspondence classifier.

(6) Key Research Accomplishments

- Collection of a database of digital mammograms for development of the CAD algorithms ------ (Task 1)

- Design a database management system for archiving the DMs, BIRADS ratings and lesion evaluation provided by radiologists on each lesion -------- (Task 1)

- Pre-processing technique for raw digital mammograms so that the CAD algorithms will be independent of the manufacturer's proprietary image processing algorithms ------- (Task 2 and Task 3)

- Comparison of Laplacian preprocessing method with GE method by evaluation of mass detection accuracy -------- (Task 2)

- Comparison of density segmentation on digitized screen-film mammograms and digital mammograms, understanding of the differences between the properties of digital mammograms and digitized mammograms ------- (Task 2, Task 3, Task 5)

- Computer aided diagnosis system for mass detection: comparison of performance on digital mammograms and digitized mammograms ------- (Task 2)

- Comparison of density weighted contrast enhancement filter and adaptive ring filter for enhancement of masses as the first step in CAD algorithm, evaluation of mass detection performance by FROC analysis ------ (Task 2, Task 6)

- Joint two-view information for computerized microcalcification detection ------ (Task 3, Task 4, Task 6)

(7) Reportable Outcomes

As a result of the support by the BCRP grant, we have conducted studies in CAD for mammography and published the results. The publications in this project year are listed in the following.

Conference Proceedings:

1. Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Zhou C. Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study. Proc SPIE 5032; 2003 (in press).

2. Hadjiiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. ROC study: Effects of computer-aided diagnosis on radiologists' characterization of malignant and benign breast masses in temporal pairs of mammograms. Proc SPIE 5032; 2003 (in press).

Conference Presentation:

1. Zhou C, Chan HP, Sahiner B, Hadjiiski LM, Paramagul C, Petrick N. Computer-aided diagnosis on mammograms using multiple image analysis: nipple identification for registration of multiple views. Presentation at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, December 1-6, 2002. Radiology 2002; 225(P): 645.

2. Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Zhou C. Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2003.

3. Hadjiiski LM, Chan HP, Sahiner B, Helvie MA, Blane C, Paramagul C, Petrick N, Roubidoux MA, Bailey J, Klein K, Foster M, Patterson S, Adler D. ROC Study: Effects of Computer-Aided Diagnosis on Radiologists' Characterization of Malignant and Benign Breast Masses in Temporal Pairs of Mammograms. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2003.

4. Hadjiiski LM, Helvie MA, Sahiner B, Chan HP, Roubidoux MA, Nees A, Patterson S, Blane C, Paramagul C, Bailey J, Klein K, Foster M, Adler D, Shen J. ROC Study: Effects of Computer-Aided Diagnosis on Radiologists' Characterization of Malignant and Benign Breast Masses in Two View Temporal Pairs of Mammograms. Submitted for presentation at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003.

5. Chan HP, Wei, J, Zhou C, Helvie MA, Roubidoux MA, Bailey J, Hadjiiski LM, Sahiner B. Comparison of mammographic density estimated on digital mammograms and screen-film mammograms. Submitted for presentation at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003.

6. Hadjiiski LM, Chan HP, Sahiner B, Zhou C, Helvie MA, Roubidoux MA. Computerized Regional Registration of Corresponding Masses and Microcalcification Clusters on Temporal Pairs of Mammograms for Interval Change Analysis. Submitted for presentation at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003.

7. Sahiner B, Chan HP, Hadjiiski LM, Helvie MA, Roubidoux MA, Petrick N. Computerized detection of microcalcifications on mammograms: Improved detection accuracy by combining features extracted from two mammographic views. Submitted for presentation at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003.

8. Petrick N, Chan HP, Sahiner B, Helvie MA, Hadjiiski LM. Evaluation of CAD Mass Detection on Prior Mammograms Containing Breast Cancers. Submitted for presentation at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003.

9. Wei, J, Sahiner B, Chan HP, Petrick N, Hadjiiski LM, Helvie MA. Computer Aided Diagnosis System for Mass Detection: Comparison of performance on Full-Field Digital Mammograms and

Digitized film Mammograms. Submitted for presentation at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003.

10. Zhou C, Hadjiiski LM, Sahiner B, Chan HP, Helvie MA, Wei, J. Computerized mammographic breast density estimation: Expectation-Maximization estimation and neural network classification of breast density. Submitted for poster presentation at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003.

(8) Conclusions

Under the support of this grant, we have investigated various computer-aided diagnosis (CAD) methods for detection of masses and microcalcifications on mammograms. We first collected a data set of full field digital mammograms that contain mammographic lesions from our breast imaging division in the Department of Radiology. The images include the manufacturer's processed images and unprocessed (raw) images. We have also developed a database management program to store the coded case information to facilitate archiving and retrieval of the cases.

To reduce the dependence of our CAD system on the manufacturer's proprietary image preprocessing method, we use the raw image as the input to our CAD system. The raw image is first preprocessed with an in-house developed Laplacian pyramid image enhancement technique. We designed the processing parameters so that the image appearance is matched to the corresponding image processed by the GE's proprietary preprocessing method. To verify the image quality of the processed image, we evaluated the mass detection accuracy of our CAD algorithm when applied to our processed images and the GE's processed images. We found that the detection sensitivities on the two sets of images were within a few percent over the entire FP range of interest. This result confirms that our preprocessing method is similar to the GE method. After this baseline is established, we can adjust the preprocessing parameters to optimize the CAD system without depending on the manufacturer's algorithm. We believe that this will allow us to develop a CAD system that can be adapted to other FFDM systems easily.

We also compared the performance of the mass detection algorithms on digitized screen-film mammograms and direct digital mammograms. After adjustment of the processing parameters in the algorithm, the detection accuracies of the algorithms on both sets of images are comparable. This confirms that the CAD system that was developed for digitized screen-film mammograms can be adapted to direct digital mammograms.

We therefore started the adaptation of the mass detection algorithm to the digital mammograms. We first investigated the effects of the image enhancement filter on the accuracy of mass detection. By replacing the DWCE filter with an adaptive ring filter, we found that the sensitivity of mass detection can be improved up to 15 %. This provides a foundation upon which we can further improve the CAD system for digital mammograms by optimization of the various steps in the detection process.

We also compared the mammographic density segmented from digitized film mammograms and direct digital mammograms. Using a data set that contained both types of the images from the same patients, we found that the correlation of the segmented breast density between the two types of images is very high. However, the estimated percent dense area on digital mammograms is, on average, about 5% lower than that estimated from digitized film mammograms. This difference may lead to improved sensitivity for lesion detection on digital mammograms.

Two-view information fusion method is being developed for correlating the detected lesions on the two views of mammograms, similar to radiologists' approach for mammographic interpretation. We found that the detection accuracy for microcalcifications can be improved by fusing of information from two mammographic views. This result demonstrates the usefulness of our proposed two-view fusion methods. We will continue to improve the two-view fusion technique and apply it to both microcalcification and mass detection.

In conclusion, we have investigated a number of areas in CAD of mammographic lesions. We have made progress in the six tasks proposed in the project. This lays the strong foundation for us to continue the development of the CAD system for digital mammograms in the coming years.

(9) References

1. Dippel S, Stahl M, Wiemker R, Blaffert T. Multiscale contrast enhancement for radiographies: Laplacian pyramid versus fast wavelet transform. IEEE Trans. Med. Img. 2002; 21: 343-353.

2. Burt PJ,Adelson EH. The Laplacian pyramid as a compact image code. IEEE Transactions on Communications 1983; COM-31: 337-345.

3. Vuylsteke P,Schoeters E. Multiscale image contrast amplification (MUSICATM). Proc SPIE 1994; 2167: 551-560.

4. Stahl M, Aach T, Buzug TM, Dippel S, Neitzel U. Noise-resistant weak-structure enhancement for digital radiography. Proc SPIE 1999; 3661: 1406-1417.

5. Wei J, Hagihara Y, Kobatake H. Detection of Cancerous Tumors on Chest X-ray Images: Candidate Detection Filter and Its Application. Proc. ICIP, Kobe, Japan, 1999; 26: 1-5.

6. Paquerault S, Petrick N, Chan HP, Sahiner B, Helvie MA. Improvement of computerized mass detection on mammograms: Fusion of two-view information. Medical Physics 2002; 29: 238-247.

7. Sahiner B, Petrick N, Chan HP, Paquerault S, Helvie MA, Hadjiiski LM. Recognition of lesion correspondence on two mammographic views - A new method of false-positive reduction for computerized mass detection. Proc. SPIE 2001; 4322: 649-655.

8. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer P. The digital database for screening mammography. In: Yaffe MJ, ed. Digital Mammography; IWDM 2000. Toronto, Canada, Medical Physics Publishing, 2001; 457-460.

(10) Appendix

Copies of the following publications are enclosed with this report.

Conference Proceedings:

1. Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Zhou C. Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study. Proc SPIE 5032; 2003 (in press).

2. Hadjiiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. ROC study: Effects of computer-aided diagnosis on radiologists' characterization of malignant and benign breast masses in temporal pairs of mammograms. Proc SPIE 5032; 2003 (in press).

Design of Three-Class Classifiers in Computer-Aided Diagnosis: Monte Carlo Simulation Study

Heang-Ping Chan*, Berkman Sahiner, Lubomir M. Hadjiiski, Nicholas Petrick[a], Chuan Zhou
Department of Radiology, The University of Michigan, Ann Arbor, MI 48109
[a]Center for Devices and Radiological Health, U.S. Food and Drug Administration, Rockville, MD 20857

ABSTRACT

For the development of computer-aided diagnosis (CAD) systems, a classifier that can effectively differentiate more than two classes is often needed. For example, a detected object on an image may need to be classified as a malignant lesion, a benign lesion, or normal tissue. Currently, a three-class problem is usually treated as a two-stage, two-class problem, in which the detected object is first differentiated as a lesion or normal tissue, and, in the second stage, the lesion is further classified as malignant or benign. In this work, we explored methods for classification of an object into one of the three classes, and compared the three-class approach with the common two-class approach. We conducted Monte Carlo simulation studies to evaluate the dependence of the performance of 3-class classification schemes on design sample size and feature space configurations. A k-dimensional multivariate normal feature space with three classes having different means was assumed. Linear classifiers and artificial neural networks (ANNs) were examined. ROC analysis for the 3-class approach was explored under simplifying conditions. A performance index representing the normalized volume under the ROC surface (NVUS) was defined. Linear classifiers for classification of three classes and two classes were compared. We found that a 3-class approach with a linear classifier can achieve a higher NVUS than that of a 2-class approach. We further compared the performance of an ANN having three or one output nodes with a linear classifier. At large sample sizes, a 3-output-node ANN was basically the same as that of a one-output-node ANN. When the three class distributions had equal covariance matrices and the distances between pairs of class means were equal, the linear classifiers could reach a higher performance for the test samples than the ANN when the design sample size was small; the linear classifier and the ANNs approached the same performance in the limit of large design sample size. However, under complex feature space configurations such as the class means located along a line, the class in the middle was poorly differentiated from the other two classes by the linear classifiers for any dimensionality; the ANN outperformed the linear classifier at all design sample size studied. This simulation study may provide some useful information to guide the design of 3-class classifiers for various CAD applications.

KEY WORDS: Computer-aided diagnosis, classifier design, 3-class classification, linear classifier, artificial neural networks, Monte Carlo simulation, likelihood ratio, ROC analysis

1. INTRODUCTION

For the development of computer-aided diagnosis (CAD) systems, a classifier that can effectively differentiate more than two classes is often needed. For example, in an automated lesion detection and characterization system, it will be important to differentiate malignant lesions from benign lesions and normal tissue. A common approach is to treat this as a two-stage classification problem having two classes at each stage; masses are distinguished from normal tissue in the first stage, and then are classified as malignant and benign in the second stage. Alternatively, if the main interest is to detect only malignant lesions, a two-class classifier is trained to differentiate the malignant class from the combined class of the other two. The two classes that are treated as one may have very different characteristics and the classification may not be optimal if the classifier is forced to recognize their features as the same. For certain types of classification tasks, a properly designed 3-class classifier can be more effective in distinguishing one class from the other two classes. The design of 3-class classifiers has not been investigated systematically in the CAD area. In this work, we performed a simulation study to explore some properties of the 3-class and 2-class classification schemes.

* chanhp@umich.edu

2. MATERIALS AND METHODS

For an m-class classification problem in which a feature vector, **x**, is to be classified into one of m classes, a common approach is to apply the Bayes' rule to minimize the misclassification rate[1]. To accomplish this, the posterior probability of **x** belonging to class i is estimated as

$$p(c_i|\mathbf{x}) = g\,P(c_i)\,p(\mathbf{x}|c_i), \qquad \text{for } i=1, \ldots, m \tag{1}$$

where $P(c_i)$ is the prior probability of class c_i, $p(\mathbf{x}|c_i)$ is the probability density of **x** in class c_i, and g is a constant. The feature vector is then assigned to class k, where k denotes the class that **x** has the maximum posterior probability,

$$k = \arg \max_{i=1,\ldots m} \{p(c_i \mid \mathbf{x})\} \tag{2}$$

However, it is difficult to estimate the posterior probability when the sample size is small. Furthermore, the misclassification rate does not take into account the fact that different types of misclassifications or correct classifications have different costs or utilities. A more general formulation of the m-class problem assigns a utility for each correct and incorrect decision, and optimizes the expected utility. The optimal decision rule depends on the utilities, as well as the prior probabilities of the classes. Let P_{Ij} denote the probability of deciding class c_i when the true class is c_j, and U_{Ij} denote the utility of deciding class c_i when the true class is c_j. The optimal decision rule is the one that maximizes the expected utility, which can be written as

$$E\{utility\} = \sum_{I=1}^{m}\sum_{j=1}^{m} U_{Ij}P_{Ij}P(c_j) \tag{3}$$

The limitation with the classifiers that maximize the correct classification rate or maximize the expected utility with fixed U_{Ij}'s is that they do not cover the entire range of sensitivity and specificity for the classification task. A receiver operating characteristic (ROC) analysis will provide the entire range of operating points. However, a 3-class classification problem will require a six-dimensional (6-D) ROC analysis as follows. For a 3-class problem with classes a (malignant), b (benign), and n (normal), there are nine possible "decision-truth" Ij pairs and hence nine probabilities: P_{Aa}, P_{Ba}, P_{Na}, P_{Ab}, P_{Bb}, P_{Nb}, P_{An}, P_{Bn}, P_{Nn}. Since the sum of every three of these probabilities is unity, e.g., $P_{Aa} + P_{Ba} + P_{Na} = 1$, only six of the nine probabilities are independent. Therefore, the ROC analysis will include these six possible variables.

For the 3-class problem, it has been shown that three decision lines that depend on two likelihood ratios (LRs) will provide the optimal decision boundaries on the LR plane as shown in Fig. 1[2] (C. E. Metz, private communication). The likelihood ratio between classes i and j is defined as the ratio of the probability density of **x** under each class,

$$LR_{ij}(\mathbf{x}) = p(\mathbf{x}|c_i)/\,p(\mathbf{x}|c_j) \tag{4}$$

In Fig. 1, the two LRs are chosen to be LR_{na} and LR_{ba}. The slopes and intercepts of the decision lines in the likelihood ratio plane depend on the prior probabilities of the classes, as well as the utilities of the different types of decisions, U_{Aa}, U_{Ba}, U_{Na}, U_{Ab}, U_{Bb}, U_{Nb}, U_{An}, U_{Bn}, U_{Nn}. The three decision lines always intersect at a common point. Varying the utilities and the priors over their allowed ranges will move the decision lines over the LR plane. For each configuration of the decision lines, the six probabilities can be estimated, producing a point in the 6-D ROC space. The complete treatment of a 6-D ROC analysis is therefore very complicated and has not yet been dealt with. In this study, we attempted to explore some properties of a 3-class problem under simplifying conditions.

We assume that the utilities can take on values in [0,1]. For correct decisions, the utilities will have the maximum value of 1, i.e., $U_{Aa} = U_{Bb} = U_{Nn} = 1$. If a malignant case is misdiagnosed as normal or benign, the utilities will be at a minimum of 0, $U_{Ba} = U_{Na} = 0$. If a normal case is called benign or vice versa, it may not be very harmful or costly so that the utilities $U_{Nb} = U_{Bn} = 1$. If a normal case or a benign case is called malignant, it will involve additional diagnostic tests or treatment and also cause patient anxiety or morbidity, the utilities U_{Ab} and U_{An} will be somewhere between 0 and 1. Under our assumptions that U_{Ab} and U_{An} are variable in (0,1) and the rest of the utilities are fixed as described above, it can be shown that two of the decision lines are reduced to one (Fig. 2), the third decision line becomes indeterminate, and the expected utility of the classification task in Eq. (3) depends only on three of the probabilities, P_{Aa}, P_{Ab}, and P_{An}. The 6-D ROC analysis will therefore be reduced to a 3-D ROC analysis under these conditions. An example of the 3-D ROC surface is shown in Fig. 3. Note that P_{Aa} is the true-positive fraction (TPF) or the sensitivity, P_{Ab} is the false-positive fraction from the benign class (FPF_b), and P_{An} is the false-positive fraction from the normal class (FPF_n). This 3-D ROC surface is therefore similar to the commonly used 2-D ROC curve except that the FPF is split into the benign and normal classes. In analogy with the 2-D ROC analysis, we can define a performance index as the normalized volume under the 3-D ROC surface (NVUS) given by

$$\text{Normalized volume under 3D ROC surface (NVUS)} = \frac{\text{Volume under 3D ROC surface}}{\text{Projected area on the FP plane}} \qquad (5)$$

Note that the NVUS can be interpreted as the average sensitivity over the range of FPF of interest, similar to the area under the 2-D ROC curve.

Ideally, if the feature vectors are transformed onto the LR plane, one can vary the decision line and determine the samples that fall into the region that is decided to be class A. The probabilities P_{Aa}, P_{Ab}, and P_{An} can then be estimated and the 3-D ROC surface generated. However, when the sample size is small, it is difficult to estimate the probability densities and derive the LRs for each **x**.

It is well-known that for the two-class classification problem in a k-dimensional feature space, the linear discriminant analysis projects the k-D feature space onto a 1-D decision axis. The decision boundary is then a threshold chosen along the decision axis. If the two class distributions are multivariate normal with equal covariance matrices, the linear discriminant classifier corresponds to the LR classifier and is optimal. This approach may be generalized to an m-class problem in a k-D feature space. In this case, the k-D feature space is projected to an (m-1)-D decision space, the decision boundaries are formed by (m-1) boundaries in the decision space[3]. For a 3-class problem (m=3), the k-D feature space is projected to a 2-D decision plane and the decision boundaries can be formed by two lines on the plane. In general, this projection is not optimal because it is not equivalent to a projection onto the LR plane. If the three class distributions are multivariate normal with equal covariance matrices, the linear transformation to a 2-D decision plane can be shown to be equivalent to a transformation to the log-likelihood ratio, *ln*(LR), plane and optimal decision boundaries can be formed on this plane.

In this preliminary study, we studied the 3-class classification problem by linearly projecting the k-D feature space to the 2-D decision plane and used two linear decision boundaries for differentiating the malignant class from the benign and the malignant classes. The classification performance was evaluated in the 3-D ROC space as shown in Fig. 3.

The 3-class classification was compared to the approach of treating the benign and normal classes as one (b+n) class such that the differentiation of the malignant class (class a) from the (b+n) class was considered to be a 2-class classification problem. The k-D feature space was thus projected to the 1-D decision line by linear discriminant analysis. This is equivalent to forming a hyperplane in the k-D feature space to separate class a from class (b+n).

We further assumed a simple k-D feature space in which the class distributions were multivariate normal, the covariance matrices for classes a, b, n were described by I, αI, αI, respectively, where I was the identity matrix and α was a constant. The mean vectors for the three classes were located at the vertices of an equilateral triangle. These characteristics are invariant upon projection to the 2-D decision plane in the 3-class classification approach described above although the scales may be changed. The 2-D decision plane in the 3-class classification approach shown in Fig. 4(a) and the example of the feature space in 2-D shown in Fig. 4(b) therefore have similar appearances. The symmetry of the class distributions about the vertical axis simplifies our analysis that follows, but the approaches should be applicable to non-symmetrical feature spaces.

For the 3-class classification approach, the slopes and intercepts of the linear decision boundaries were varied over the entire plane. For each set of boundaries, we could calculate the three probabilities, P_{Aa}, P_{Ab}, and P_{An} and generate a point in the 3-D ROC space. The surface formed by the highest sensitivity (P_{Aa}) at each FP location corresponded to the best decision boundaries. The NVUS was then derived from the highest sensitivity surface relative to its projected area on the FP plane.

For the 2-class classification approach with linear discriminant analysis, the best projection of the decision axis would be parallel to the symmetry (vertical) axis because of the symmetry of the class distributions. The decision boundary along this axis thus corresponded to a hyperplane perpendicular to the symmetry line. The decision boundary is illustrated as a horizontal line in the 2-D feature space (Fig. 4(b)). By moving the decision boundary along the decision axis and scoring the TPF and FPF, we could generate the 2-D ROC curve and derive the area under the ROC curve, A_z.

We compared the 3-class and 2-class approaches in two different ways. First, we compared the area under ROC curve under similar situations. For the 3-class approach and in our feature space with symmetry, the slice of the 3-D ROC surface along the diagonal of $P_{Ab} = P_{An}$ was equivalent to the situation of treating class b and class n equally, i.e., $U_{Ab} = U_{An}$. We calculated the area under the ROC curve obtained from this slice and compared it with the A_z obtained in the 2-class approach. In the second comparison, we modified the 2-class classification approach in the original k-D feature space. If we allowed the hyperplane to orient at an angle to the symmetry axis (the best projected decision axis in the linear discriminant analysis), it was similar to taking into consideration that there were different utilities of making FP decisions from class b or class n. For example, if the slope of the decision boundary was positive as shown in Fig. 5(a), we were less concerned with deciding a class-b sample as class a than deciding a class-n sample as class a so that it implied $U_{Ab} > U_{An}$. On the other hand, if the slope of the decision boundary was negative as shown in Fig. 5(b), we were less concerned with deciding a class-n sample as class a than deciding a class-b sample as class a so that it implied $U_{Ab} < U_{An}$. Therefore, by varying the slope and intercept of the single decision boundary in the 2-class approach, we could also generate a 3-D ROC surface and calculate its NVUS. We then compared the NVUS obtained from the 3-class and 2-class approaches.

Neural Network Classifiers

Another common approach that is often applied to the m-class classification problem is to use an artificial neural network (ANN) classifier with (m-1) output nodes. During training, the desired output of a sample from the i^{th} class is assigned 1 at the i^{th} node and assigned 0 at all other nodes. Under ideal conditions (sufficiently large training sample size and proper training), it has been shown that the ANN approaches a Bayes' classifier and the output for a given sample at the i^{th} node approaches the posterior probability of the sample in the i^{th} class[4]. Therefore, a properly trained ANN can be used for transforming the feature space to the LR plane and the 6-D ROC analysis applied. However, since the available design sample size is often limited in practice, the training of an ANN is usually far from being ideal. One of the common methods of analyzing the ANN output is to apply a 2-D ROC analysis to the scores of an individual output node, e.g., the i^{th} node, to distinguish the i^{th} class from the other classes. In this study, we evaluated the application of ANNs having one output node and three output nodes to the three-class problem. For training of the ANN with one output node, the desired output of the class-a samples was assigned to be 1 and those of the class-b and class-n samples were assigned to be 0. This is equivalent to treating the classification task as a 2-class problem without distinction between class b and class n. For training of the ANN with three output nodes, the desired output of a sample from the i^{th} class ($i = 1, 2, 3$) was assigned to be 1 at the i^{th} node and 0 at all other nodes. Under ideal conditions, one of the output nodes is actually redundant because the output of the third node is complementary to the other two. For both the 1-output-node ANN and the 3-output-node ANN, we applied 2-D ROC analysis to the output node that distinguished class a from the other two classes and compared the A_z values.

Simulation Study

We performed a simulation study to evaluate the different approaches discussed above. For this study, we assumed that the three class distributions were multivariate normal with diagonal covariance matrices. In a given experiment, 1000 samples were randomly drawn from the population for each of the classes. A subset of N_{train} trainers was randomly drawn from the 1000 available samples of each class and the rest, ($1000-N_{train}$), of the samples were held out as testers. N_{train} was varied from 20 to 600 per class for the linear classification study, and varied from 20 to 500 for the ANN study. For each condition, the experiment was repeated 50 times such that a new set of 1000 samples per class were drawn from the population. The dependence of the performance index, either A_z or NVUS, for each of the classification approaches on training sample size was evaluated. The ANNs were assumed to have one hidden layer with the number of hidden nodes equal to the number of input nodes. Backpropagation with a delta-bar-delta rule was used for training of the ANNs.

3. RESULTS

For comparison of the 3-class and 2-class approaches using linear classification, we assumed a 12-D multivariate normal feature space with covariance matrices *I*, 8*I*, 8*I* for class a, b, n, respectively. The comparison of A_z as a function of $1/N_{train}$ is plotted in Fig. 6. It can be seen that, for a given approach, when the design sample size is limited, the training (resubstitution) A_z is optimistically biased and the test (holdout) A_z is pessimistically biased, in comparison with the A_z at $N_{train} \rightarrow \infty$. The biases decrease as N_{train} increases. In the limit of $N_{train} \rightarrow \infty$, the training and test A_z approach essentially the same value. The A_z obtained from the 3-class approach is consistently higher than that from the 2-class approach for a given N_{train}.

Fig. 7 shows the comparison of the NVUS for the 3-class and 2-class approaches using linear classification in the same feature space. The characteristics of the curves are very similar to those observed in Fig. 6. The training NVUS is optimistically biased whereas the test NVUS is pessimistically biased compared to the limit achieved with large design sample size. The NVUS from the 3-class approach is again consistently higher than that from the 2-class approach for a given N_{train}.

For the comparison of the 3-output-node and 1-output-node ANNs, we first assumed a k-D (k=3, 6, 9, 12) multivariate normal feature space with equal covariance matrices *I*, *I*, *I* for class a, b, n, respectively. The dependence of A_z on $1/N_{train}$ is shown in Fig. 8(a) for the 3-output-node ANN and in Fig. 8(b) for the 1-output-node ANN. The characteristics of the A_z-versus-$1/N_{train}$ curves are very similar to those obtained in our previous study of 2-class classification problems[5]. The training A_z is optimistically biased and the test A_z is pessimistically biased compared with the A_z values at $N_{train} \rightarrow \infty$. The biases increase with the dimensionality of the feature space for a given N_{train} and decrease with increasing design sample size. It can be seen that the A_z values in the limit of $N_{train} \rightarrow \infty$ are very similar for the 3-output-node and the 1-output-node ANNs. For a given N_{train}, the biases of the 3-output-node ANN are larger than those of the 1-output-node ANN for the high dimensional feature spaces, probably because of the larger number of weights that need to be trained in the 3-output-node ANN with the finite design samples. For comparison, we also trained a linear discriminant classifier to differentiate class a from class (b+n) and plotted the A_z–versus-$1/N_{train}$ curves in Fig. 8(c). The A_z values in the limit of $N_{train} \rightarrow \infty$ from the linear classifiers are again very similar to those from the ANNs. These results indicate that the 3-output-node or the 1-output-node ANNs is basically performing 2-class classification at each of its output nodes. It is interesting to note that, when N_{train} is small, the biases in A_z from the linear classifier are much smaller than those from the ANNs. Therefore, in this feature space, when the design sample size is small, a linear classifier may be preferred over the ANNs because the performance of the trained linear classifiers is superior to that of the ANNs for unknown test samples.

The relative performance of the ANNs and linear classifiers depends strongly on the configuration of the class distributions, however. This can be demonstrated by comparing their performances in another multivariate normal feature space with unequal covariance matrices: class a had an identity matrix *I*, class b had a diagonal matrix with its diagonal elements varying from 1 to 2 in equal increment, class n had a diagonal matrix with its diagonal elements varying from 1 to 3 in equal increment. The three class means were lined up along a straight line in the k-D feature space. Fig. 9 shows an example of the class distributions in a 2-D feature space. The performances of the three classifiers in distinguishing class a, which is in the middle, from class b and class n are compared in Figs. 10(a) to 10(c). Under these conditions, the 3-output-node classifiers had slightly higher test A_z when N_{train} was small, but the A_z in the limit of $N_{train} \rightarrow \infty$ seemed to approach a level slightly lower than those of the 1-output-node ANN for the higher dimensional (9-D and 12-D) feature spaces. As expected, the linear classifiers were not able to distinguish the class a in the middle of class b and class n. Their performance was close to random guess for all sample sizes. These indicated that ANNs can be superior to a linear classifier for classification tasks with complex class distributions.

4. CONCLUSIONS

In this study, we explored some properties of 3-class and 2-class approaches to a 3-class classification task under simplifying conditions. By using Monte Carlo simulation study, we have examined the dependence of the performances of different classification schemes on design sample sizes for some feature space configurations. We found that a 3-class approach can achieve higher classification accuracy than a 2-class approach under some conditions. Applying a 2-D ROC analysis to the output of a 3-output-node ANN achieved similar classification accuracy as that of a 1-output-node ANN. The ANNs may not be the method of choice for some classification tasks when the available design sample size is small. A complete treatment of 3-class classification using a 6-D ROC analysis is very complex and was not attempted in this preliminary study. Further investigation is underway to investigate if 3-class approaches can improve the accuracy for some classification tasks in CAD.

ACKNOWLEDGMENTS

This work is supported by USPHS grant CA95317 and U. S. Army Medical Research and Materiel Command Grants DAMD 17-02-1-0214. The content of this publication does not necessarily reflect the position of the funding agency, and no official endorsement of any equipment and product of any companies mentioned in this publication should be inferred. The authors are grateful to C. E. Metz, Ph.D., for his helpful discussion and notes on optimal decision variable.

REFERENCES

1. M. Nadler and E. P. Smith, *Pattern Recognition Engineering*, (John Wiley and Sons, New York, 1993).
2. H. L. Van Trees, *Detection, estimation, and modulation theory*, (John Wiley and Sons, New York, 1968).
3. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, (Wiley, New York, 1973).
4. C. M. Bishop, *Neural Networks for Pattern Recognition*, (Clarendon Press, Oxford, 1995).
5. H. P. Chan, B. Sahiner, R. F. Wagner and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," Medical Physics 26, 2654-2668 (1999).

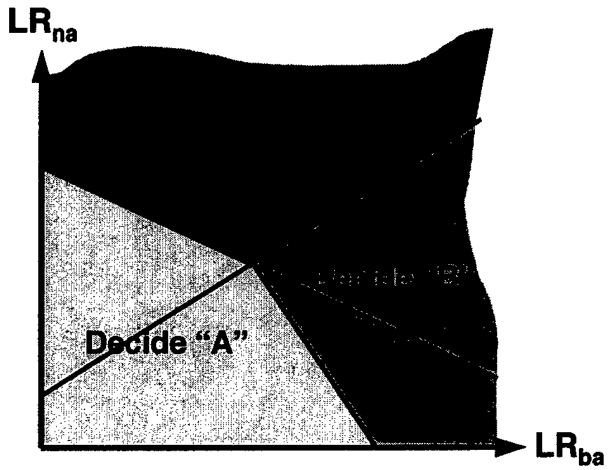


Fig. 1. Likelihood Ratio (LR) plane for a 3-class classification task.

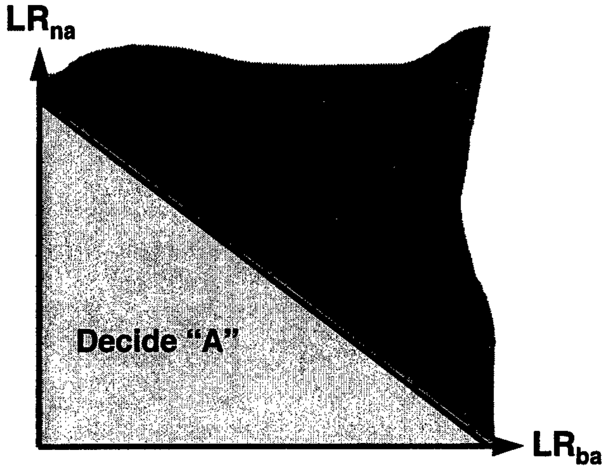


Fig. 2. Likelihood Ratio plane for a 3-class classification task under the assumptions in this study.

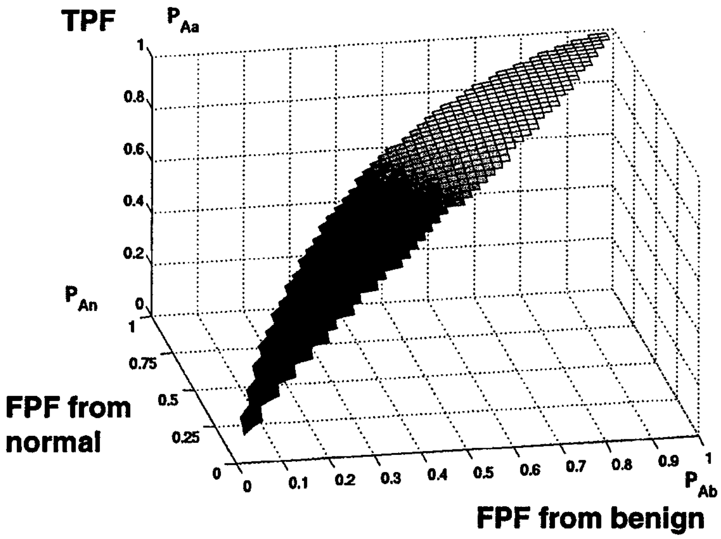


Fig. 3. 3-D ROC surface for the analysis of the 3-class classification problem under the assumptions in this study.

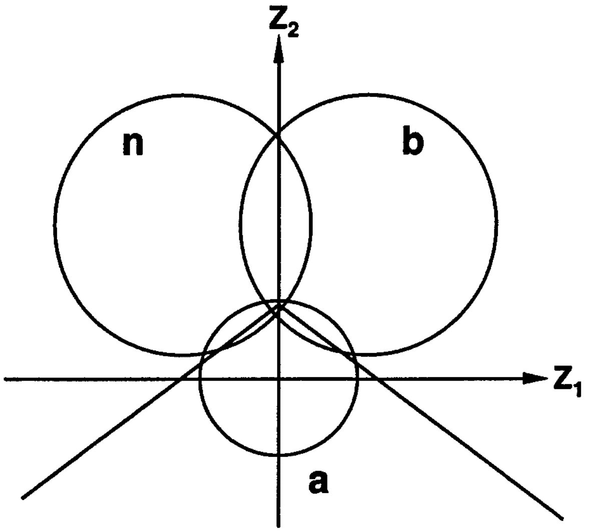


Fig. 4(a). Three-class approach for a 3-class classification task: 2-D decision plane with two linear decision boundaries.

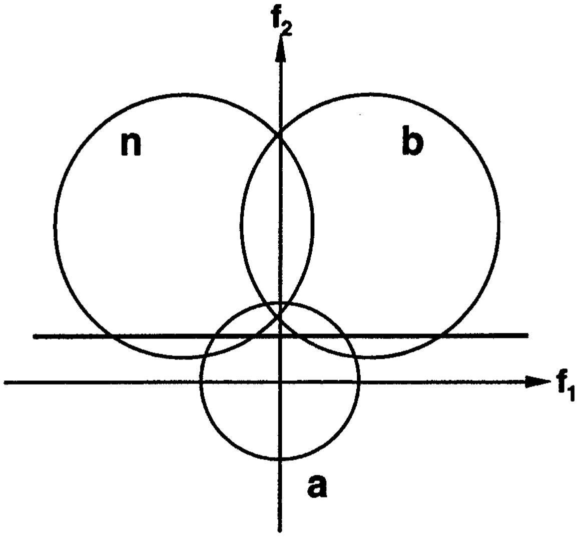


Fig. 4(b). Two-class approach for a 3-class classification task: k-D feature space (shown in 2-D as an example) with one linear decision boundary.

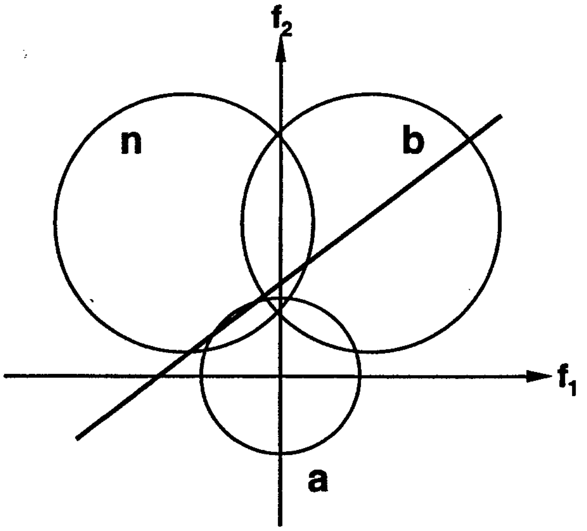


Fig. 5(a). Two-class approach for a 3-class classification task: a linear decision boundary that assumes $U_{Ab} > U_{An}$.

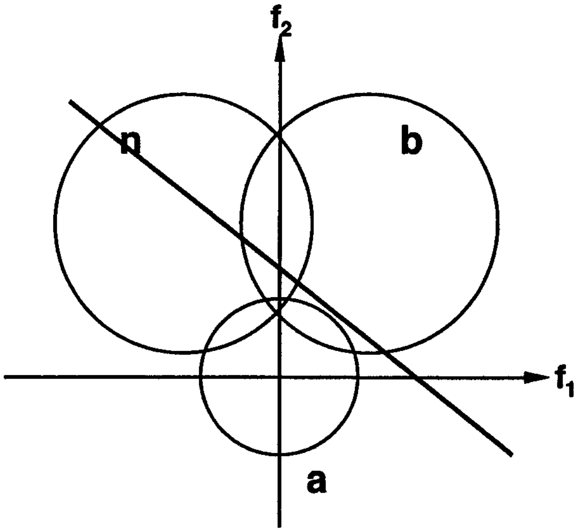


Fig. 5(b). Two-class approach for a 3-class classification task: a linear decision boundary that assumes $U_{Ab} < U_{An}$.

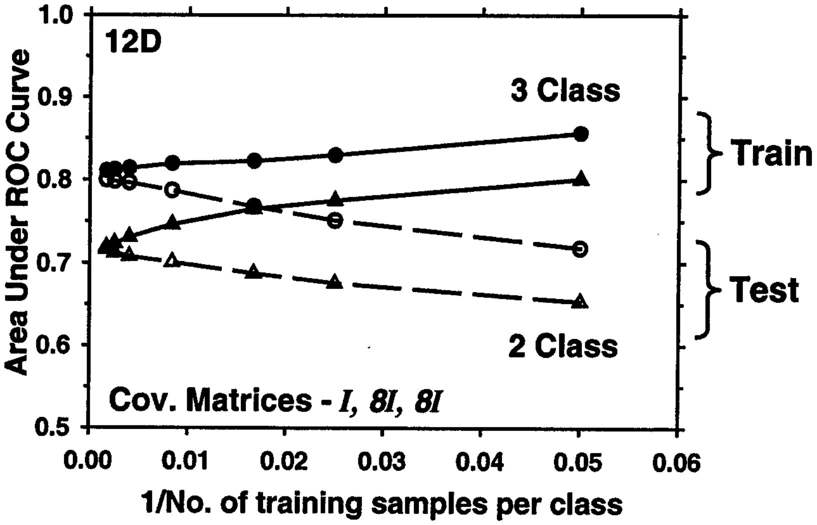


Fig. 6. Comparison of the performance of the 3-class and 2-class approaches for a 3-class problem. The areas under the 2-D ROC curves corresponding to $U_{Ab} = U_{An}$ are compared as a function of design sample size. Circles: 3-class approach. Triangles: 2-class approach. Solid curves: training results. Dashed curves: test results.

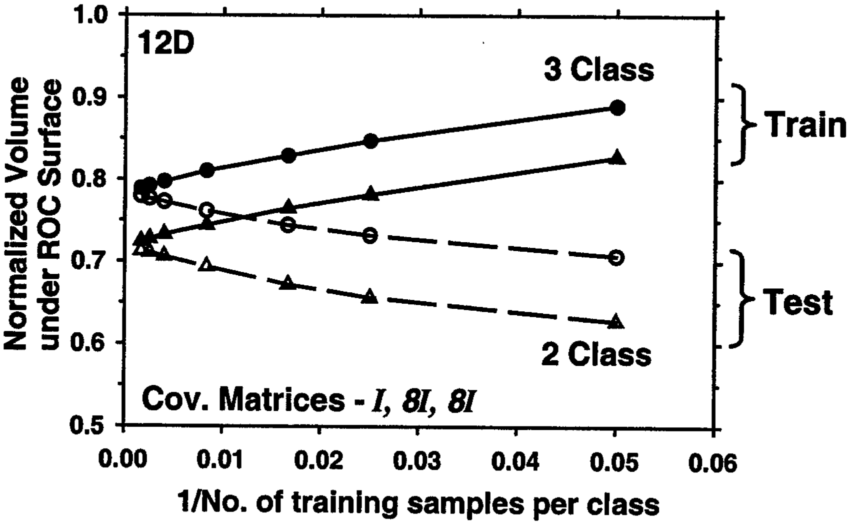


Fig. 7. Comparison of the performance of the 3-class and 2-class approaches for a 3-class problem. The normalized volumes under the 3-D ROC surface (NVUS) are compared as a function of design sample size. Circles: 3-class approach. Triangles: 2-class approach. Solid curves: training results. Dashed curves: test results.

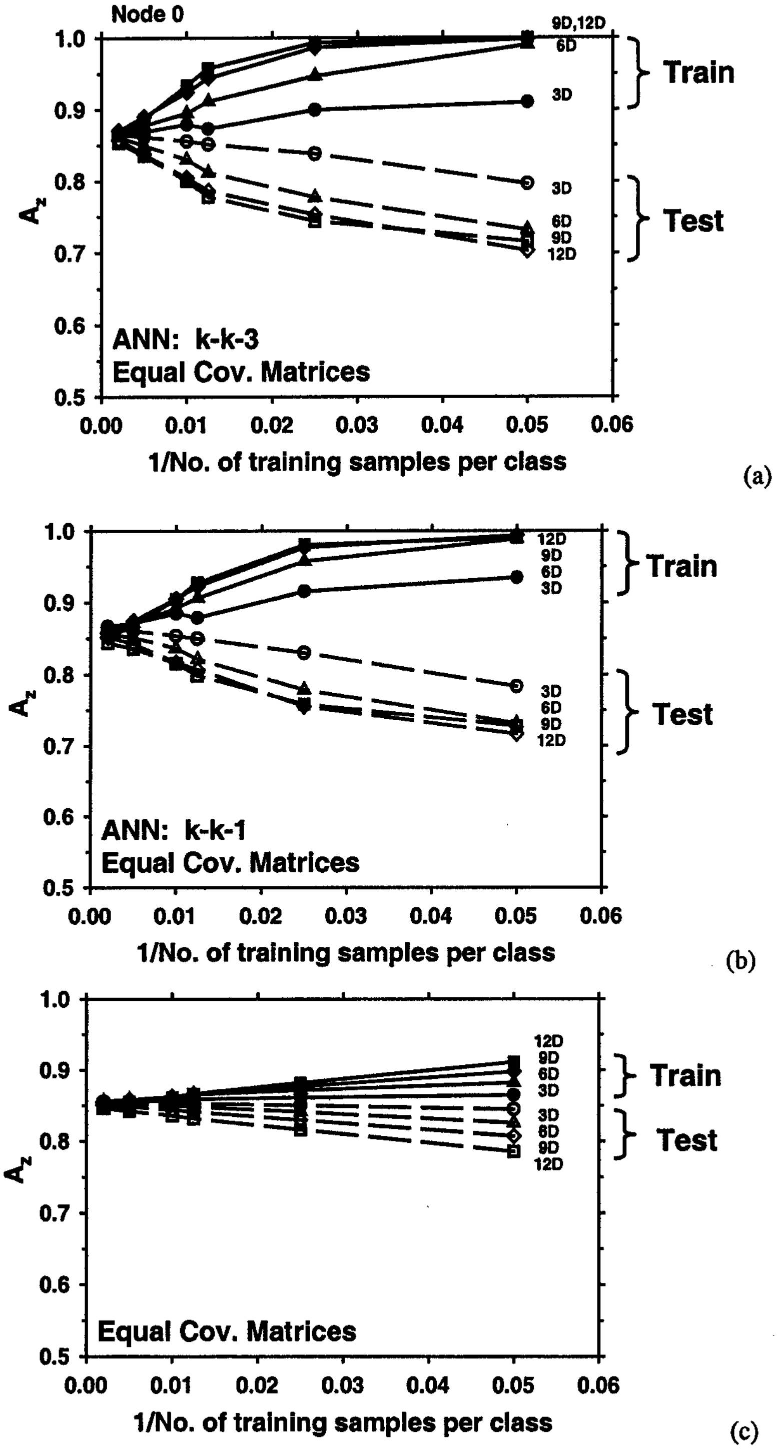


Fig. 8. Classification performance in terms of A_z for differentiating class a from class b and class n. The class distributions in 3-D, 6-D, 9-D, 12-D feature spaces are multivariate normal with equal covariance matrices and class means located at the vertices of an equilateral triangle. (a) ANN: k input nodes, k hidden nodes, 3 output nodes, (b) ANN: k input nodes, k hidden nodes, 1 output node, and (c) linear classifier. Solid curves: training results. Dashed curves: test results.

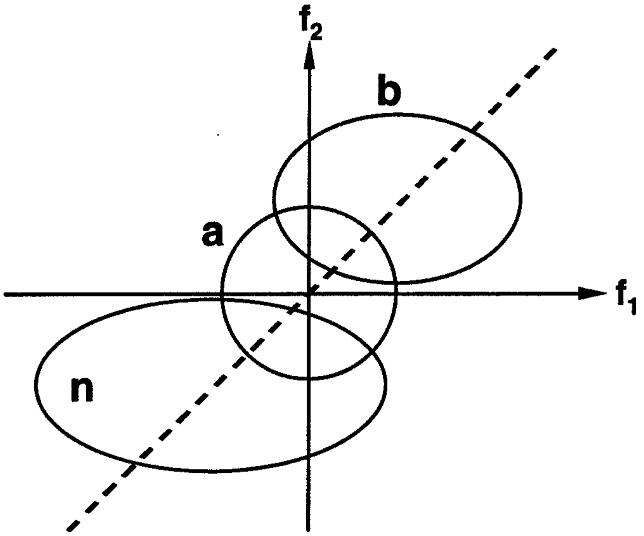


Fig. 9. A 3-class feature space with multivariate normal class distributions. The covariance matrices are diagonal and the three class means are located along a line. The example is illustrated in 2-D.

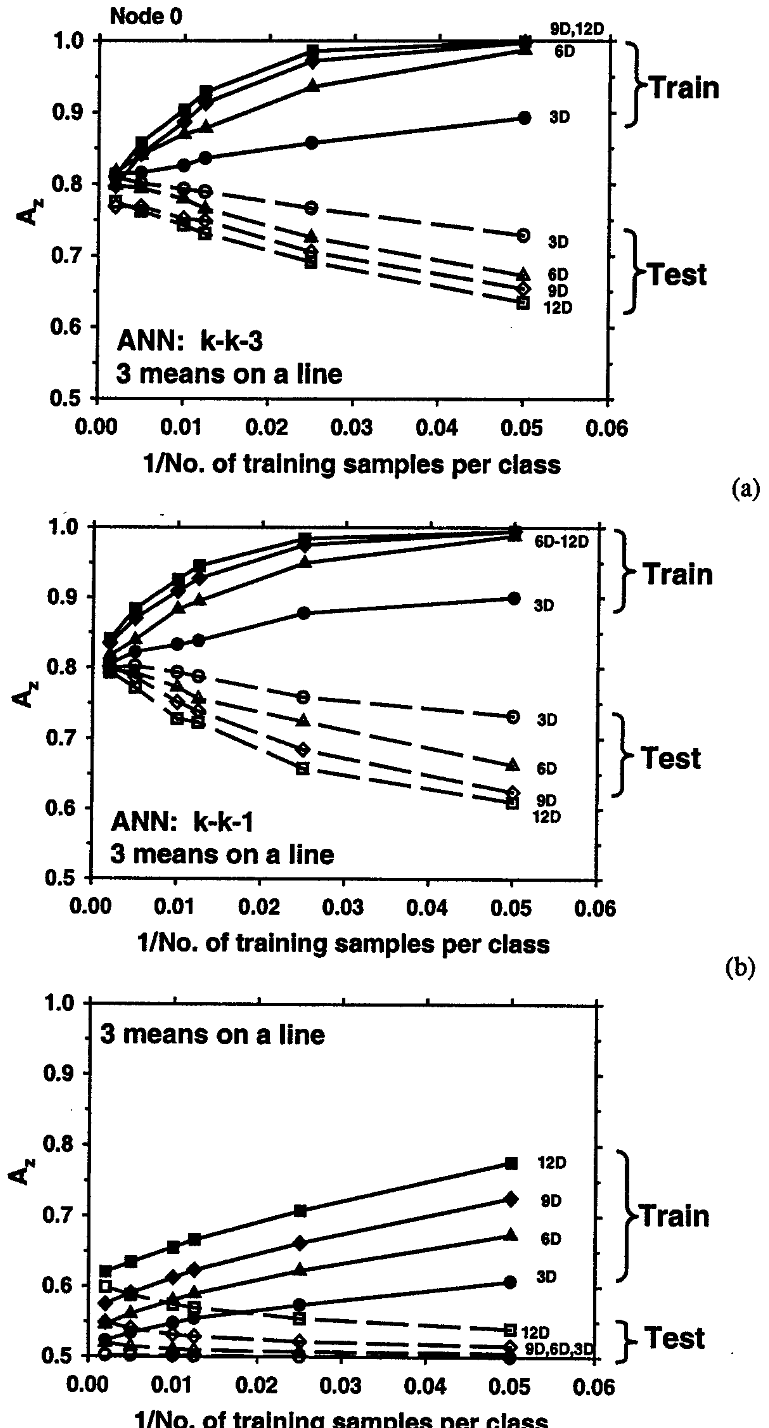


Fig. 10. Classification performance in terms of A_z for differentiating class a from class b and class n. The class distributions in 3-D, 6-D, 9-D, 12-D feature spaces are multivariate normal with unequal covariance matrices and class means along a line. (a) ANN: k input nodes, k hidden nodes, 3 output nodes, (b) ANN: k input nodes, k hidden nodes, 1 output node, and (c) linear classifier. Solid curves: training results. Dashed curves: test results.

ROC study: Effects of computer-aided diagnosis on radiologists' characterization of malignant and benign breast masses in temporal pairs of mammograms

Lubomir Hadjiiski*, Heang-Ping Chan, Berkman Sahiner, Mark A. Helvie,
Marilyn Roubidoux, Caroline Blane, Chintana Paramagul, Nicholas Petrick[a],
Janet Bailey, Katherine Klein, Michelle Foster, Stephanie Patterson,
Dorit Adler, Alexis Nees, Joseph Shen

Department of Radiology, University of Michigan, Ann Arbor, MI;
[a]Center for Devices and Radiological Health, U.S. Food and Drug Administration, Rockville, MD 20857

ABSTRACT

We conducted an observer performance study using receiver operating characteristic (ROC) methodology to evaluate the effects of computer-aided diagnosis (CAD) on radiologists' performance for characterization of masses on serial mammograms. The automated CAD system, previously developed in our laboratory, can classify masses as malignant or benign based on interval change information on serial mammograms. In this study, 126 temporal image pairs (73 malignant and 53 benign) from 52 patients containing masses on serial mammograms were used. The corresponding masses on each temporal pair were identified by an experienced radiologist and automatically segmented by the CAD program. Morphological, texture, and spiculation features of the mass on the current and the prior mammograms were extracted. The individual features and the difference between the corresponding current and prior features formed a multidimensional feature space. A subset of the most effective features that contained the current, prior, and interval change information was selected by a stepwise procedure and used as input predictor variables to a linear discriminant classifier in a leave-one-case-out training and testing resampling scheme. The linear discriminant classifier estimated the relative likelihood of malignancy of each mass. The classifier achieved a test A_z value of 0.87. For the ROC study, 4 MQSA radiologists and 1 breast imaging fellow assessed the masses on the temporal pairs and provided estimates of the likelihood of malignancy without and with CAD. The average A_z value for the likelihood of malignancy estimated by the radiologists was 0.79 without CAD and improved to 0.87 with CAD. The improvement was statistically significant (p=0.0003). This preliminary result indicated that CAD using interval change analysis can significantly improve radiologists' accuracy in classification of masses and thereby may increase the positive predictive value of mammography.

Keywords: Computer-Aided Diagnosis, Interval Changes, ROC Observer Study, Classification, Mammography, Breast Cancer.

1. INTRODUCTION

Mammography is currently the most sensitive method for detecting early breast cancer, and it is also the most practical screening exam [1,2] compared with other breast imaging techniques. However, the specificity of mammography is relatively low, only 15-30% of suspected breast lesions recommended for biopsy are actually malignant [3-5]. The unnecessary biopsies increase health care costs and cause patient anxiety and morbidity. If the specificity of differentiating malignant and benign mammographic lesions can be improved, the efficacy of mammography will be enhanced.

* L. H. (correspondence): e-mail:lhadjiski@umich.edu

One of the important techniques that radiologists use in mammographic interpretation is to compare the current mammograms of a patient with those obtained in previous years, if available. The interval change information can help the detection of abnormalities, and identification of malignant breast lesions. It is shown that comparison with prior mammograms can improve both the sensitivity and specificity in breast cancer diagnosis [6,7].

In an early investigation, Chan et al. [8] demonstrated that computer-aided diagnosis (CAD) could improve significantly radiologists' detection of subtle mammographic microcalcification in an ROC study, This promising result stimulated continued development of CAD systems. To date, a number of CAD algorithms have been developed to detect suspicious masses and microcalcifications and to distinguish malignant and benign lesions on mammograms. Several ROC studies have shown that CAD systems could improve radiologists' accuracy in characterization of breast lesions. It has also been reported that CAD systems can increase the detection of breast cancers on screening mammograms in clinical practice[9,10].

Chan et al [11] performed an observer study to evaluate the effects of CAD, designed for characterization of malignant and benign masses on single view mammograms[12], on radiologists' diagnostic accuracy. They found that the radiologists' accuracy for classification of masses as malignant or benign in terms of the area under receiver operating characteristic (ROC) curve (A_z) was significantly improved ($p=0.022$ for one-view reading and 0.007 for two-view reading) with CAD compared to that without CAD. Huo et al [13] also conducted an observer study with 12 radiologists to classify masses on multiple views of mammograms. They also found that the radiologists' performance in terms of A_z was significantly improved ($p=0.001$) with computer aid. Jiang et al [14] performed an observer study to evaluate the effect of CAD on radiologists' classification of microcalcification clusters on mammograms. They found that with the computer aid the radiologists achieved a statistically significant improvement ($p<0.0001$).

The CAD systems for lesion classification so far employed information from a single exam.[12,14-19]. Based on the experiences of radiologists, it can be expected that even higher accuracy may be achieved if the computer can utilize the interval change information from multiple exams for classification. We recently[20] developed a classification scheme that combines prior and current information automatically extracted from masses on prior and current mammograms, respectively. We found that the classifier using the combined prior and current information performed significantly better ($p=0.015$) in terms of A_z than the classifier using current information alone. The current study investigated the effects of CAD on assisting radiologists in evaluating interval changes in serial mammograms. To our knowledge, this is the first ROC experiment to evaluate the effects of a computer classifier using interval change information on radiologists' diagnosis of breast cancers.

2. MATERIALS AND METHODS

2.1 Data set

We selected a set of 126 temporal pairs of mammograms containing biopsy-proven masses on the current mammograms from our database. The mammograms in the database were digitized consecutively from the patients who had undergone breast biopsy in our department. The selection criterion used in the current study was that the case had serial exams in which a corresponding mass could be identified. The mammograms thus contained masses covering a range of sizes and conspicuity that will be seen in clinical practice. The data set consisted of 220 mammograms from 52 patients. The mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of $50\mu m \times 50\mu m$ and 4096 gray levels. The image matrix size was reduced by averaging every 2 x 2 adjacent pixels and down-sampled by a factor of 2 to obtain images with a pixel size of $100\mu m \times 100\mu m$ for analysis by the computer.

There were 53 biopsy proven masses (32 malignant and 21 benign) in the 52 cases. The 220 mammograms contained different mammographic views (CC, MLO, and lateral views) and multiple serial examinations of the masses including the examination when the biopsy decision was made. By matching masses of the same view from two different examinations, a total of 126 temporal pairs were formed, of which 73 were malignant and 53 benign. Since all cases in this data set had undergone biopsy, the benign masses in this set could not be distinguished easily from the malignant ones based on current mammographic criteria. For the malignant masses in this data set, the average mass size was 7.9 mm on the prior mammograms and 12.0 mm on the current mammograms. The corresponding sizes were 9.8 mm and 11.4 mm, respectively, for the benign masses.

To simulate a more realistic clinical situation 34 additional temporal pairs containing corresponding normal structures in the serial mammograms were also included. In this way the radiologist also has to distinguish mass-mimicking fibroglandular tissue from malignant masses. The temporal pairs had a time interval of 6 to 48 months. More than 67% of the pairs had a time interval of 12 months.

2.2 Design of classifier for classification of masses in serial mammograms

We have developed a novel classification technique that utilizes the current and prior information on serial mammograms to characterize the masses. The classification technique has been described in detail elsewhere[20]. The method is summarized in the flowchart shown in Figure 1. Initially a region of interest (ROI) containing the mass was defined by a radiologist on both the current and prior mammograms. Automatic segmentation of the mass within each ROI was performed based on an active contour model [21,22]. A set of texture, morphological, and spiculation features were extracted for each mass.

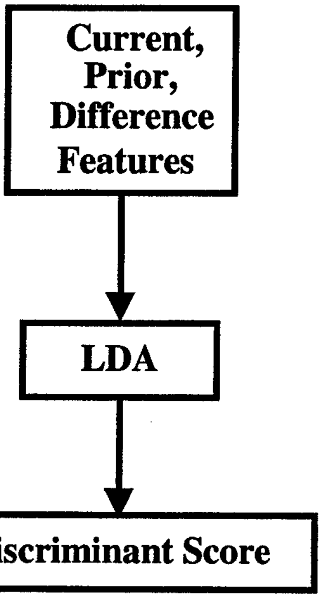


Figure 1. Block-diagram of the classification method.

The texture features were based on run-length statistics (RLS) matrices [23]. The RLS matrices were computed from the images obtained by the rubber band straightening transform (RBST) [12]. The RBST maps a band of pixels surrounding the mass onto a rectangular region. Five texture measures were extracted from the vertical and horizontal gradient images derived from the RBST image in two directions [12]. Therefore, for each ROI, a total of 20 RLS features were calculated. Morphological features were extracted from the automatically segmented mass shape and gray levels [22,24]. Spiculation features were extracted by using the statistics of the image gradient direction relative to the normal direction to the mass border in a ring of pixels surrounding the mass [21,22]. A total of 35 features (20 RLS, 12 morphological and 3 spiculation) were therefore extracted from each ROI. Additionally, difference features were obtained by subtracting a prior feature from the corresponding current feature, resulting in 35 difference features.

A "leave-one-case-out" resampling scheme was used for the training and testing of the classifier. In order to reduce the dimensionality of the feature space, a stepwise feature selection was employed to select the most effective feature subset from each training cycle. An average of 7 features were selected for the classification task from the training subsets.

A relative malignancy rating by the computer classifier on a scale of 1 to 10 was provided to the radiologists for the reading with CAD. The relative malignancy rating was obtained by linearly scaling the classifier output within the interval between 1 and 10 and then rounding the result to the closest integer. A higher rating corresponded to a higher likelihood of being malignant. Gaussian functions were fitted to the distributions of the malignant and benign samples to obtain a fitted binormal distribution with the classifier's malignancy ratings scaled to the range of 1 to 10 (Figure 2). The fitted distribution was displayed on the graphical user interface as a reference when the radiologist evaluated the cases using CAD.

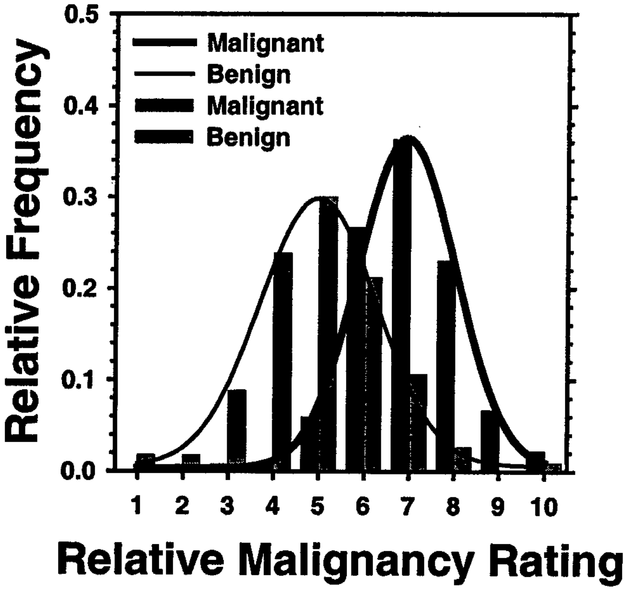


Figure 2. Binormal distribution fitted to the histogram.

2.3 Radiologist's classification of masses in serial mammograms

The observer study was designed to compare radiologists' performance on the classification of malignant and benign breast masses with and without CAD. The ROIs extracted from the current and the prior mammograms containing the corresponding mass was displayed side-by-side on a display monitor. The observers' performance was evaluated under two reading conditions. In the first reading condition, the radiologist read the temporal image pair of the mass without computer aid. In the second reading condition, the radiologist read the temporal pair with computer classifier's relative malignancy rating of the mass displayed on the screen. The observer was asked to provide an estimate of the likelihood of malignancy of the mass in a 100-point rating scale under each reading condition. Four MQSA radiologists and one breast imaging fellows participated as observers in this study.

A counter-balanced design was used in arranging the reading orders in different modes and the case orders in different reading sessions for the observers. This approach would minimize the potential effects such as learning, fatigue, and memorization on the outcomes of the observer experiments. A graphic user interface was developed for the purpose of presenting the temporal pairs of mass ROIs to the radiologists and recording their ratings. Each observer underwent a training session before the actual reading sessions to familiarize them with the performance of the CAD system and the experimental procedure.

2.4 ROC analysis

The likelihood of malignancy ratings of the individual observers for the two reading conditions were analyzed by using ROC methodology. A binormal ROC curve was fitted to each observer's 100-point rating data for each reading condition by the LABROC program using maximum likelihood estimation.[25] The classification accuracy was quantified by using the total area under the fitted ROC curve, A_z,. The slope and the intercept parameters for the individual ROC curves were also estimated by the LABROC program. For each reading condition, the average performance of the radiologists was estimated as the area under an average ROC curve, which was derived from the average slope and intercept parameters of the 5 individual observer's ROC curves for that reading condition. The statistical significance of the difference in A_z between the two reading conditions was estimated by the Student's two-tailed paired t-test on the 5 pairs of individual observer's A_z values.

3. RESULTS

The A_z values for the 5 radiologists participating in the study for the two reading conditions with and without CAD are presented in Fig 3. The computer classifier's test A_z value was 0.87. The average ROC curves for the 5 observers when reading with and without CAD were plotted in Fig.4. The A_z value from the average ROC curve was 0.79 for reading without CAD and 0.87 for reading with CAD. The radiologist performance was improved, both individually and on average, when reading with the CAD system. The improvement in the average A_z between the reading without CAD and the reading with CAD was statistically significant (Student's two-tailed paired t-test, p=0.0003).

The computer classifier's A_z value of 0.87 was higher than the individual radiologists' A_z values obtained under the reading condition without CAD. The relatively low accuracy of the radiologists in classifying the masses reflected the fact that these were difficult cases that all had been recommended for biopsy. All five radiologists improved their accuracy in classification of the malignant and benign masses when the CAD system was available as a second opinion. Two radiologists achieved an A_z higher than that of the computer classifier under the reading condition with CAD. We did not observe specific differences between the breast imaging fellow compared to the MQSA-approved radiologists. The improvement in A_z ranged between 0.06 and 0.1.

4. CONCLUSION

We have performed an observer ROC study to evaluate the effects of computer-aided diagnosis on radiologists' characterization of masses on serial mammograms. In this observer study the radiologists improved their performance with statistically significance (p = 0.0003) when their reading without computer aid was compared to that with computer aid. These results suggest that CAD may be helpful in improving the accuracy of malignant and benign mass characterization.

ACKNOWLEDGMENTS

This work was supported by USAMRMC grants DAMD17-98-1-8211, DAMD17-02-1-0489, and DAMD17-02-1-0214. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC program.

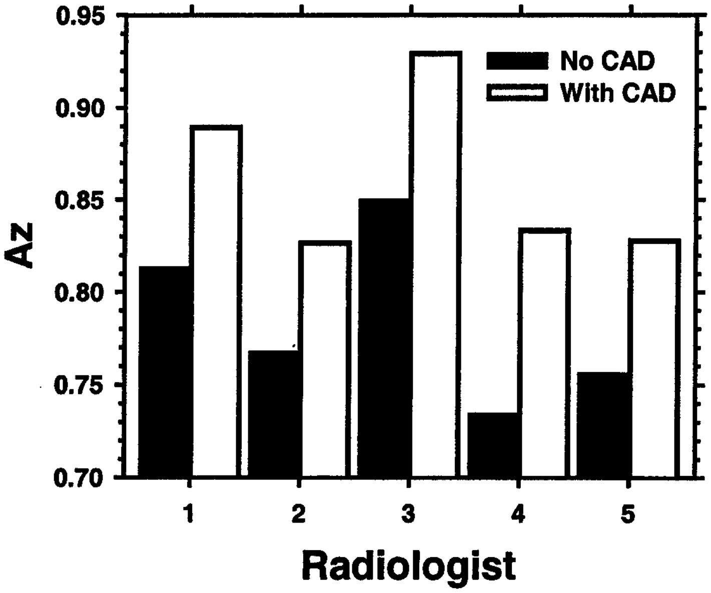


Figure 3. The area under ROC curve, A_z, for the characterization of the masses in 126 pairs of serial mammograms by 5 radiologists under two reading conditions: without CAD and with CAD. The average A_z for the two reading conditions: no CAD (A_z=0.79), with CAD (A_z=0.87).

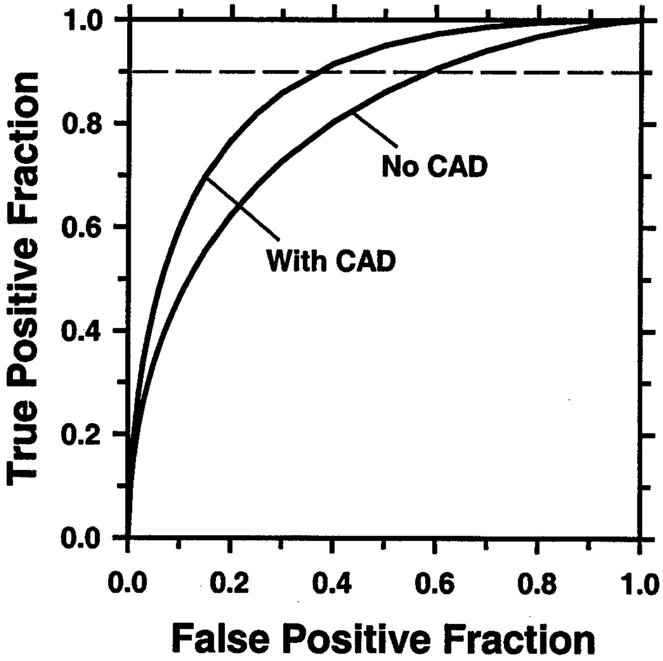


Figure 4. Area under ROC curve for the mode without CAD and the mode with CAD by the 5 radiologists. Average area for the two reading modes: No CAD (A_z=0.79), With CAD (A_z=0.87). The difference is statistically significant (Student paired t-test, p=0.0003).

REFERENCES

1. H. C. Zuckerman, "The role of mammography in the diagnosis of breast cancer," *In: Breast Cancer, Diagnosis and Treatment*, 152-172, Eds. I. M. Ariel and J. B. Cleary, McGraw-Hill, New York, 1987.

2. L. Tabar and P. B. Dean, "The Control of Breast Cancer through Mammography Screening," *Radiologic Clincs of North America* **25**, 961, 1987.

3. E. A. Sickles, "Mammographic features of 300 consecutive nonpalpable breast cancers," *American Journal of Roentgenology* **146**, 661-663, 1986.

4. D. B. Kopans, "The positive predictive value of mammography," *American Journal of Roentgenology* **158**, 521-526, 1991.

5. D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," *Current Opinion in Radiology* **4**, 123-129, 1992.

6. L. W. Bassett, B. Shayestehfar, and I. Hirbawi, "Obtaining previous mammograms for comparison: usefullness and costs," *Amer. J. Roentgenology* **163**, 1083-1086, 1994.

7. E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: results in 3183 consecutive cases," *Radiology* **179**, 463-468, 1991.

8. H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," *Investigative Radiology* **25**, 1102-1110, 1990.

9. M. A. Helvie, L. M. Hadjiiski, E. Makariou, H. P. Chan, N. Petrick, and S. B. Lo, "A Non-Commercial CAD System for Breast Cancer Detection on Screening Mammograms Achieves High Sensitivity : A Pilot Clinical Trial," *Radiology* **225(P)**, 459, 2002.

10. T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," *Radiology* **220**, 781-786, 2001.

11. H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," *Radiology* **212**, 817-827, 1999.

12. B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Medical Physics* **25**, 516-526, 1998.

13. Z. M. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast Cancer: Effectiveness of Computer-aided Diagnosis - Observer Study with Independent Database of Mammograms," *Radiology* **224**, 560-568, 2002.

14. Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Academic Radiology* **6**, 22-33, 1999.

15. H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Leung, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," *Physics in Medicine and Biology* **42**, 549-567, 1997.

16. Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Academic Radiology* **5**, 155-168, 1998.

17. J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," *IEEE Transactions on Medical Imaging* **12**, 664-669, 1993.

18. L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, and M. A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," *IEEE Transactions on Medical Imaging* **18**, 1178-1187, 1999.

19. G. D. Tourassi, M. K. Markey, J. Y. Lo, and C. E. Floyd, "A neural network approach to breast cancer diagnosis as a constraint satisfaction problem," *Medical Physics* **28**, 804-811, 2001.

20. L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. N. Gurcan, "Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses," *Medical Physics* **28**, 2309-2317, 2001.

21. B. Sahiner, H. P. Chan, N. Petrick, L. M. Hadjiiski, M. A. Helvie, and S. Paquerault, "Active contour models for segmentation and characterization of mammographic masses," *The 5th International Workshop on Digital Mammography*, 357-362, Toronto, Canada, 2001.

22. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics* **28**, 1455-1465, 2001.

23. M. M. Galloway, "Texture classification using gray level run lengths," *Computer Graphics and Image Processing* **4**, 172-179, 1975.

24. N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," *Medical Physics* **26**, 1642-1654, 1999.

25. D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "ROC rating analysis: Generalization to the population of readers and cases with the jackknife method," *Investigative Radiology* **27**, 723-731, 1992.